

A Sinkhorn-NN Hybrid Algorithm for Optimal Transport

Jonathan Geuter

October 15, 2022

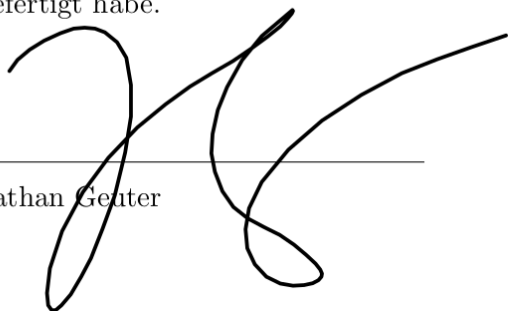
Master's Thesis

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 Hilfsmittel angefertigt habe.

A handwritten signature in black ink, consisting of a large, stylized 'J' followed by a series of loops and a long horizontal stroke extending to the right.

Jonathan Genter

15.10.2022, Berlin

Datum, Ort

Abstract

The Sinkhorn algorithm [11] is the state-of-the-art to compute approximations to optimal transport distances between discrete probability distributions, using an entropic regularizer added to the optimal transport problem. The entropic problem being a strictly 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 starting vectors like the n -dimensional one-vector 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/SinkhornNNHybrid>.

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}, μ) und (\mathcal{Y}, ν) mit ihren jeweiligen Borel- σ -Algebren und eine Kostenfunktion $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ lautet das *Kantorovich-Problem*

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

wobei $\Pi(\mu, \nu)$ die Menge aller *Transportpläne* enthält, das heißt aller Wahrscheinlichkeitsmaße γ auf dem Produktraum $\mathcal{X} \times \mathcal{Y}$, sodass $\gamma \circ \pi_{\mathcal{X}}^{-1} = \mu$, $\gamma \circ \pi_{\mathcal{Y}}^{-1} = \nu$. Ein Maß $\gamma \in \Pi(\mu, \nu)$ "transportiert" somit das Maß μ auf ν , und über alle Produktmaße γ mit dieser Eigenschaft werden die Gesamtkosten minimiert. Das duale Problem lautet

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

und beide Probleme besitzen stets Optimierer, deren Optima übereinstimmen. In der Praxis ist es jedoch insbesondere in hohen Dimensionen sehr rechenaufwändig, diese Optimierer zu bestimmen. Ein effizienter Algorithmus zur Approximation des Optimums und auch eines optimalen Transportplans im Fall, dass beide Maße diskret sind, ist der Sinkhornalgorithmus [11]. Dieser konvergiert gegen die Lösung des entropisch regularisierten Problems

$$\inf_{\gamma \in \Pi(\mu, \nu)} \{ \langle c, \gamma \rangle - \varepsilon H(\gamma) \},$$

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 (vgl. Amos et al. [2] oder auch Thornton und Cuturi [40]). Wir zeigen, dass eine gut gewählte Initialisierung die Konvergenzgeschwindigkeit deutlich 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

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

gegeben μ und ν 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, der sogenannten *Wassersteindistanzen* und des Sinkhornalgorithmus (Kapitel 3 und 4). Im Anschluss werden die Details unseres Sinkhorn-NN-Algorithmus erläutert (Kapitel 5) und Ergebnisse verschiedener Experimente präsentiert (Kapitel 6), die abschließend diskutiert werden (Kapitel 7). Notationen werden in Kapitel 2 erklärt, und einige technische Details und Hintergründe lassen sich im Appendix A finden.

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	22
3.5	Duality Theorem	26
3.6	Wasserstein Distances	28
3.7	Discrete Optimal Transport	32
4	Sinkhorn Algorithm	35
4.1	Entropic Optimal Transport	35
4.2	Sinkhorn Algorithm	41
4.3	Initializing Sinkhorn's Algorithm	43
5	Sinkhorn-NN Hybrid Algorithm	45
5.1	A Trained Initialization for the Sinkhorn Algorithm	45
5.2	Training Data	46
5.3	Test Data	48
5.4	Network Architecture	49
5.5	Why Not...?	50
5.6	Training	54
6	Results	55
6.1	Error w.r.t. Iterations	55
6.2	Speed	56
7	Discussion	61
A	Appendix	62
	References	67
	Index	70

1 Introduction

Optimal Transport [41][33][35] is, in short, the theory of optimally transporting something from one place to another. Today, it plays an increasing role in various areas. Besides economics [21], it thrives in machine learning applications, amongst others, and has been used in domain adaptation [10], single-cell genomics [36], imitation learning [12], imaging [37] and signal processing [26], to name a few.

French mathematician Gaspard Monge laid its foundation in the 18th century. In his 1781 publication *Mémoire sur la théorie des déblais et des remblais* [29], 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}, μ) and (\mathcal{Y}, ν) , equipped with their Borel- σ -algebras, and a cost function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$, the task is to find a measurable *transport map* $T : \mathcal{X} \rightarrow \mathcal{Y}$ 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

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

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 [24] in 1942. 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 have to find 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

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

This problem has a dual problem:

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

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 $\mathcal{X} = \{x_1, \dots, x_m\}$ and $\mathcal{Y} = \{y_1, \dots, y_n\}$ are both finite, an efficient way to compute an approximation of the solution is the Sinkhorn algorithm [11], an iterative algorithm converging to the solution of the *entropic optimal transport problem*, which consists of adding an entropic regularizer to the Kantorovich problem:

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

where $\varepsilon > 0$ is a regularizing coefficient and $H(\gamma) := -\sum_{ij} \gamma_{ij} (\log \gamma_{ij} - 1)$ the entropy of γ . At its core, the algorithm consists of initializing $v^0 \in \mathbb{R}_{>0}^n$ and the simple iterates

$$u^{l+1} = \frac{\mu}{\exp(-c/\varepsilon)v^l}, \quad v^{l+1} = \frac{\nu}{\exp(-c/\varepsilon)^{\top} u^{l+1}}, \quad l = 0, 1, 2, \dots,$$

where the fractions are to be understood as element-wise division and we slightly abuse notation by considering $\mu \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^n$ to be vectors representing the measures. As the entropic optimal transport problem can be shown to be ε -strongly convex it admits a unique solution, and the Sinkhorn algorithm is guaranteed to converge to this solution. However, carefully choosing a starting vector v^0 for the algorithm can significantly improve its convergence speed. In the literature, little attention has been paid to the initialization of the Sinkhorn algorithm so far (see, e.g., Amos et al. [2] or Thornton and Cuturi [40]). 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

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

given μ and ν . This potential can be used to compute a starting vector v^0 for the Sinkhorn algorithm via

$$v^0 = \exp(f^c/\varepsilon),$$

where f^c is the c -transform of f , defined as $f^c(x) = \min_y c(x, y) - f(y)$. We will show 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. Additionally, 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 and further explanations omitted during the thesis.

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. Some basic definitions first:

- $\mathbb{N} := \{0, 1, 2, 3, \dots\}$
- $\mathbb{N}_{>0} := \{1, 2, 3, 4, \dots\}$
- for $m, n \in \mathbb{N}_{>0}$, $m \leq n$: $\llbracket m, n \rrbracket := \{m, m+1, \dots, n\}$ and $\llbracket n \rrbracket := \llbracket 1, n \rrbracket$
- 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 $\llbracket n \rrbracket$
- $2^{\mathcal{X}} := \{X : X \subset \mathcal{X}\}$ for a set \mathcal{X}

Linear Algebra

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. Similarly, 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 $\Delta^{n-1} = \{v \in \mathbb{R}_{\geq 0}^n : \sum_i v_i = 1\}$ be the $n - 1$ -dimensional probability simplex in \mathbb{R}^n and $\Delta_{>0}^{n-1} = \{v \in \Delta^{n-1} : v_i > 0 \text{ for all } i \in \llbracket n \rrbracket\}$. 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. We will also write $[a_{ij}]_{ij}$ for the matrix A . By $\text{vec}(A)$ we refer to the vector in \mathbb{R}^{mn} that one gets by concatenating the columns of A , i.e. $\text{vec}(A)_{(j-1)m+i} = A_{ij}$ for all i, j . 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

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

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a function, then by $f(A)$ we refer to the matrix resulting from entry-wise application of f , i.e.

$$f(A) := [f(a_{ij})]_{ij} \in \mathbb{R}^{m \times n}.$$

For two vectors $a, b \in \mathbb{R}^n$, $\langle a, b \rangle$ is the usual scalar product and $\text{diag}(a) \in \mathbb{R}^{n \times n}$ the matrix with diagonal entries $\text{diag}(a)_{ii} = a_i$ 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*:

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

Analysis

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

A function $f : \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is called *lower semicontinuous* if

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

The support $\text{supp}(f)$ of a function $f : \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is the set $\overline{\{x \in \mathcal{X} : f(x) \neq 0\}}$.

By $C(\mathcal{X})$ we denote the space of all continuous functions $f : \mathcal{X} \rightarrow \mathbb{R}$, and the space of all continuous and bounded functions is denoted by $C_b(\mathcal{X})$.

Sometimes, for spaces \mathcal{X} and \mathcal{Y} , we will consider functions $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, $g : \mathcal{X} \rightarrow \mathbb{R}$, $h : \mathcal{Y} \rightarrow \mathbb{R}$ and then make statements like $f \leq g + h$. This is to be understood as $f(x, y) \leq g(x) + h(y)$ for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$.

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 π_X is the projection $X \times Y \rightarrow X$, $(x, y) \mapsto x$.

Measure Theory

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 $A \subset \mathcal{X}$, the indicator function $\mathbb{1}_A : \mathcal{X} \rightarrow \mathbb{R}$ 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 $\delta_x : \mathcal{B}(\mathcal{X}) \rightarrow \mathbb{R}$, $\delta_x(A) = \mathbb{1}_A(x)$.

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 are always 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 $\text{supp}(\mu)$ of $\mu \in P(\mathcal{X})$ is the smallest closed set on which μ is concentrated.

¹Note 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.

²More details on the product of two Polish spaces and its σ -algebra can be found in the appendix, see e.g. remark A.11, 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.

If T is a map $\mathcal{X} \rightarrow \mathcal{Y}$ 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

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

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.19).

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 $X : \Omega \rightarrow \mathcal{S}_1$ is a random variable defined on a probability space (Ω, \mathbb{P}) , 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 $\mathcal{L}(X)$. Similarly, if \mathcal{S}_2 is another measurable space and $Y : \Omega \rightarrow \mathcal{S}_2$ is another random variable defined on (Ω, \mathbb{P}) , the pushforward measure $\mathbb{P} \circ (X, Y)^{-1}$ of $(X, Y) : \Omega \rightarrow \mathcal{S}_1 \times \mathcal{S}_2$ 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, as a is not defined point-wise, statements like this are to be understood as " μ -almost surely, we have...".

Machine Learning

By *lr* we will sometimes refer to a network's learning rate. We will oftentimes deal with the *mean squared error* which we denote by *MSE*. For a set of errors $e := \{e_1, \dots, e_N\}$ with $e_i \in \mathbb{R}$ for $i \in \llbracket N \rrbracket$, it is defined as

$$\text{MSE}(e) := \frac{\sum_i e_i^2}{N}. \quad (1)$$

Also, we will sometimes write $\text{MSE}(a, b)$ for two vectors a and b . This is to be understood as $\text{MSE}(a - b)$ as in (1).

Similarly, the L^1 error is defined as

$$L^1(e) := \frac{\sum_i |e_i|}{N},$$

and again $L^1(a, b) := L^1(a - b)$.

By *ReLU* we refer to the *rectified linear unit*, an activation function defined by

$$\text{ReLU}(x) := \max\{0, x\}, \quad x \in \mathbb{R},$$

which is applied element-wise in the case where x is a vector.

By $1eZ$ for $Z \in \mathbb{Z}$ we refer to the quantity 10^Z , e.g. $1e35$, $1e-3$, etc.

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

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 considered to be a discrete metric space. In particular sections 3.1–3.3 and 3.6 are based on [41].

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.11, 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}, μ) and (\mathcal{Y}, ν) be two probability spaces. Let $X : \Omega \rightarrow \mathcal{X}$ and $Y : \Omega \rightarrow \mathcal{Y}$ be two random variables on a probability space (Ω, \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

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

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

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

This is an immediate consequence of lemma A.19 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 $T : \mathcal{X} \rightarrow \mathcal{Y}$ 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 $T : \mathcal{X} \rightarrow \mathcal{Y}$ such that $\gamma = \mu \circ (\text{Id}, T)^{-1}$.

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

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

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

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

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* in section 3.6. 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:

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

The corresponding transport costs are:

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

and

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

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

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

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 $\gamma = \mu \circ (\text{Id}, T)^{-1}$ (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.16): 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 in lemma 3.2.5.

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

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

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

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

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

We will apply the previous lemma to the sets $\{\mu\} \subset P(\mathcal{X})$ and $\{\nu\} \subset P(\mathcal{Y})$, hence we have to prove that they are tight, which might seem intuitively obvious.

Lemma 3.2.5. *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 [32].

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 $\{a_1, a_2, \dots\}$ 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 A.14. Furthermore,

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

\square

The main idea in the proof will be to use *Weierstraß' Theorem* (A.15) on the functional

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

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. In order to prove lower semicontinuity of F , we need to establish the following well-known fact about lower semicontinuous functions.

Lemma 3.2.6. *Let (\mathcal{X}, d) be a metric space and $f : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous and bounded from below. Then there exists a sequence $(f_n)_{n \in \mathbb{N}}$ of continuous functions $f_n : \mathcal{X} \rightarrow \mathbb{R}$ 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 $f_0 \leq f_1 \leq \dots \leq f$, 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 $x \in \mathcal{X}$, $\lim_{n \rightarrow \infty} f_n(x)$ exists and $\lim_{n \rightarrow \infty} f_n(x) \leq f(x)$. To finish the proof, it suffices to show that $\lim_{n \rightarrow \infty} f_n(x) \geq f(x)$.

Without loss of generality, we may assume $l := \lim_{n \rightarrow \infty} f_n(x) < \infty$ (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

$$f_n(x) \leq f(y_n) + nd(x, y_n) < f_n(x) + \frac{1}{n}. \quad (2)$$

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 $d(x, y_n) \rightarrow 0$ as $n \rightarrow \infty$, i.e. y_n converges to x . As we have $f_n(x) + \frac{1}{n} > f(y_n)$ by equation (2), we get

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

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

Remark 3.2.7. 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 3.2.8 (Lower Semicontinuity of the Cost Functional). *Let $c : \mathcal{X} \times \mathcal{Y}$ be a lower semi-*

continuous, bounded from below cost function. Then the functional

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

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

Proof. As c is bounded from below and lower semicontinuous, by lemma 3.2.6 there exists a sequence $(c_n)_{n \in \mathbb{N}}$ of continuous $c_n : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ such that $c(x, y) = \lim_{n \rightarrow \infty} c_n(x, y)$ 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

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

where in the first step, we used the Monotone Convergence Theorem (theorem A.18), 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. \square

We are now able to prove the existence of an optimal transport plan for the Kantorovich problem 3.2.1.

Theorem 3.2.9 (Existence of an Optimal Transport Plan). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two Polish spaces and $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ 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 $\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\gamma(x, y)$.*

Proof. Let

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

We need to show that F attains its minimum on $\Pi(\mu, \nu)$. By proposition 3.2.8, 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.15). From lemma 3.2.5, 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.16), $\Pi(\mu, \nu)$ 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 $(\varphi, \psi) \in C_b(\mathcal{X}) \times C_b(\mathcal{Y})$. Then

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

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

Remark 3.2.10. The lower boundedness of c ensures that $\int_{\mathcal{X} \times \mathcal{Y}} c \, d\gamma$ 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 [41], 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 3.2.11. 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 $\int_{\mathcal{X} \times \mathcal{Y}} c \, d\mu \, d\nu < +\infty$ which yields a finite cost for at least the trivial coupling.

Remark 3.2.12. 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 3.2.13. *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) 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 $B \subsetneq A$ and $\mu(B) > 0$. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ 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 [34]. □

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 $\int_{\mathcal{X} \times \mathcal{Y}} c \, d\gamma$. 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}, μ) and (\mathcal{Y}, ν) be two probability spaces and $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ a cost function. The *dual Kantorovich problem* is defined as:

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

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

Note that for the sake of simplicity, we changed the sign of ψ in our formulation of the dual. This means that $\psi(x)$ would correspond to *what you pay 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

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

which yields

$$\inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) \right\} \geq \sup_{\substack{\psi \in L^1(\mu) \\ \varphi \in L^1(\nu)}} \left\{ \int_{\mathcal{X}} \psi(x) d\mu(x) + \int_{\mathcal{Y}} \varphi(y) d\nu(y) : \psi + \varphi \leq c \right\}. \quad (3)$$

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

$$\psi(x) = \inf_y c(x, y) - \varphi(y) \text{ for all } x \in \mathcal{X}, \quad \varphi(y) = \inf_x c(x, y) - \psi(x) \text{ for all } y \in \mathcal{Y}. \quad (4)$$

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

Definition 3.3.2 (*c*-Transform). Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$, $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{\pm\infty\}$ and $\varphi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\pm\infty\}$.

The *c*-transform $\psi^c : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$ of ψ is defined via

$$\psi^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - \psi(x).^4$$

Similarly, the *c*-transform $\varphi^c : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ 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 $\psi(x) = \varphi^c(x)$

⁴We assume that for all $y \in \mathcal{Y}$ there exists some $x \in \mathcal{X}$ such that $c(x, y) - \psi(x) < \infty$, and a similar assumption on φ .

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; (4) 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 (4), 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 (4) will not make the first equation in (4) 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 $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a function.*

A function $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ is called c -concave if $\psi \not\equiv -\infty$ and there exists a function $\varphi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$ such that $\psi = \varphi^c$.

Similarly, a function $\varphi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$ is called c -concave if $\varphi \not\equiv -\infty$ and there exists a function $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ such that $\varphi = \psi^c$.

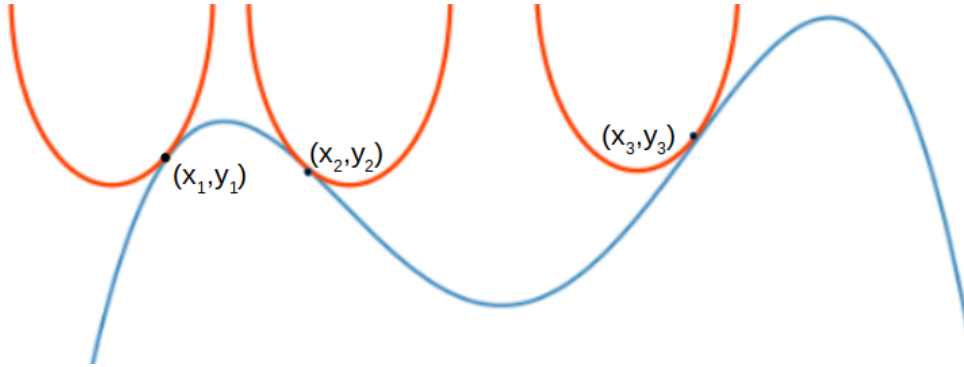


Figure 1: A c -concave function $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ is one whose graph can entirely be caressed from above with the negative cost function. The points (x_i, y_i) , $i \in \llbracket 3 \rrbracket$, come from the superdifferentials $\partial^c \psi(x_i)$ respectively, 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 $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ 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 subset of $\mathcal{X} \times \mathcal{Y}$ where $\psi^c(y) = c(x, y) - \psi(x)$ is special in the sense that on this set, the infimum from (4) is attained at $\psi(x)$ for $\varphi = \psi^c$; note that the inequality $\psi^c(y) \leq c(x, y) - \psi(x)$ holds on all of $\mathcal{X} \times \mathcal{Y}$ by definition. This set also gets a name.

Definition 3.3.5 (c -Superdifferential). *Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ and let $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ 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 $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$, 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 $\phi : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$, we have the identity $\phi^c = (\phi^{cc})^c =: \phi^{ccc}$, as

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

and $\tilde{x} = x$ yields

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

whereas $\tilde{y} = y$ yields

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

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. \square

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 $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ and let $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ 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. " \Rightarrow ": Let $y \in \partial^c \psi(x)$. We have

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

" \Leftarrow ": 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 $y \in \partial^c \psi(x)$. \square

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

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

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.

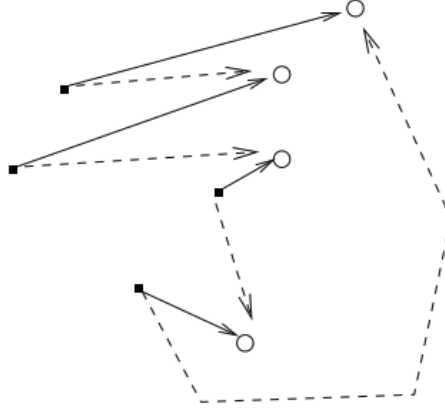


Figure 2: Trying to reduce the transport cost by finding a cycle of lower cost. Solid arrows stand for the original transport plan, dashed arrows for the rerouted mass.⁵

Definition 3.3.9 (*c*-Cyclical Monotonicity). *Let \mathcal{X} and \mathcal{Y} be two sets and $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ 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 $(x_1, y_1), \dots, (x_n, y_n) \in \Gamma$, 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 $\psi : \mathcal{X} \rightarrow \mathbb{R} \cup \{-\infty\}$ be *c*-concave. Then $\partial^c \psi$ is a *c*-cyclically monotone set.*

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

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

□

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].

⁵Source: [41], figure 5.1.

Theorem 3.4.1 (Fundamental Theorem of Optimal Transport). *Let (\mathcal{X}, μ) and (\mathcal{Y}, ν) be two Polish probability spaces and $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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) $\text{supp}(\gamma)$ 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 $\text{supp}(\gamma) \subset \partial^c \psi$

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

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

(i) \Rightarrow (ii) :

Intuitively, it is quite clear what to do: Assume that $\text{supp}(\gamma)$ 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 $\text{supp}(\gamma)$ is not c -cyclically monotone. That means we can find $n \in \mathbb{N}_{>0}$, $(x_i, y_i) \in \text{supp}(\gamma)$, $i \in \llbracket n \rrbracket$, 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

$$\sum_{i=1}^n c(u_i, v_{\sigma(i)}) - c(u_i, v_i) < 0 \quad \text{for all } (u_i, v_i) \in U_i \times V_i, \quad i \in \llbracket n \rrbracket. \quad (5)$$

Now we will construct a signed measure η (see definition A.3) 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.6), so 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) $\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\eta(x, y) < 0$ (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 $\Omega := \prod_{i=1}^n U_i \times V_i$ and $P \in \mathcal{P}(\Omega)$ be defined as $P = \gamma_1 \otimes \gamma_2 \otimes \dots \otimes \gamma_n$, where $\gamma_i := \frac{1}{m_i} \gamma|_{U_i \times V_i}$ with $m_i := \gamma(U_i \times V_i)$, $i \in \llbracket n \rrbracket$. Then each γ_i is a probability measure on $U_i \times V_i$ and P is a probability measure on Ω . Let $\pi_{U_i} : \Omega \rightarrow U_i$ and $\pi_{V_i} : \Omega \rightarrow V_i$ 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

$$\begin{aligned} \eta \circ \pi_{\mathcal{Y}}^{-1}(B) &= \eta(\mathcal{X} \times B) \\ &= \frac{\min_i m_i}{n} \sum_{i=1}^n P((U_1 \times V_1) \times \dots \times (U_{\sigma(i)} \times (V_{\sigma(i)} \cap B)) \times \dots \times (U_n \times V_n)) \\ &\quad - P((U_1 \times V_1) \times \dots \times (U_i \times (V_i \cap B)) \times \dots \times (U_n \times V_n)) \\ &= \frac{\min_i m_i}{n} \sum_{i=1}^n \gamma_{\sigma(i)}(U_{\sigma(i)} \times (V_{\sigma(i)} \cap B)) - \gamma_i(U_i \times (V_i \cap B)) = 0, \end{aligned} \quad (6)$$

hence $\eta \circ \pi_{\mathcal{Y}}^{-1} = 0$. Similarly, $\eta \circ \pi_{\mathcal{X}}^{-1} = 0$ follows, except that in the analogous derivation as in equation (6), the permutation will disappear. This proves (2).

Regarding (3), note that from (5) it follows that $\int c d\eta \leq 0$, but since c is continuous we even have $\int c d\eta < 0$.

To sum it up, we now have a transport plan $\tilde{\gamma} \in \Pi(\mu, \nu)$ such that

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

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 $(x_i, y_i) \in \Gamma$, $i \in \llbracket n \rrbracket$, there has to hold

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

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

$$\psi(x) := \inf_{\substack{n \in \mathbb{N}_{>0} \\ (x_i, y_i) \in \Gamma, i \in \llbracket n \rrbracket}} (c(x, y_1) - c(x_1, y_1)) + (c(x_1, y_2) - c(x_2, y_2)) + \dots + (c(x_n, \bar{y}) - c(\bar{x}, \bar{y})). \quad (7)$$

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

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

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

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

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

$$\psi(x) \leq c(x, \bar{y}) - c(\bar{x}, \bar{y}) \leq a(x) + b(\bar{y}) - c(\bar{x}, \bar{y}),^6$$

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 (7) we get

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

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 $\text{supp}(\gamma) \subset \partial^c \psi$. We have

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

Also $\max\{\psi^c, 0\} \in L^1(\nu)$ 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 $\max\{\psi^c, 0\} \in L^1(\nu)$. Let $\tilde{\gamma} \in \Pi(\mu, \nu)$

⁶To be precise, $b(\bar{y})$ is not really defined, as $b \in L^1(\nu)$. However, b is defined ν -a.s., and this suffices as the point $(\bar{x}, \bar{y}) \in \Gamma$ was arbitrary.

be an arbitrary transport plan. We will show that $\int c d\gamma \leq \int c d\tilde{\gamma}$, which yields the claim:

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

□

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 $\text{supp}(\tilde{\gamma}) \subset \text{supp}(\gamma)$, 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 $\text{supp}(\gamma) \subset \partial^c \psi$. Actually, an ever stronger statement holds true: For any other optimal transport plan $\tilde{\gamma}$, we have $\text{supp}(\tilde{\gamma}) \subset \partial^c \psi$ *with the same function* ψ . Indeed,

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

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

Remark 3.4.5. As we have seen in theorem 3.2.9, 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}, μ) and (\mathcal{Y}, ν) be two Polish probability spaces and $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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:

$$\inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) = \sup_{\substack{(\psi, \varphi) \in L^1(\mu) \times L^1(\nu) \\ \psi + \varphi \leq c}} \int_{\mathcal{X}} \psi(x) d\mu(x) + \int_{\mathcal{Y}} \varphi(y) d\nu(y). \quad (8)$$

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 3.2.9, 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 (8). As we have seen before, the infimum on the left hand side is greater or equal than the supremum on the right hand side, see equation (3). For the reverse inequality, let $\gamma \in \Pi(\mu, \nu)$ be optimal. By theorem 3.4.1 and its proof, we know there exists a c -concave function ψ such that $\text{supp}(\gamma) \subset \partial^c \psi$, $\max\{\psi, 0\} \in L^1(\mu)$, and $\max\{\psi^c, 0\} \in L^1(\nu)$. This gives us

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

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

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

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

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 $\text{supp}(\gamma) \subset \partial^c \psi$: By theorem 3.4.1 and its proof, 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

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

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.

Remark 3.5.4. The proof of theorem 3.5.1 shows that we not only have $\max\{\psi, 0\} \in L^1(\mu)$, but actually $\psi \in L^1(\mu)$ and $\psi^c \in L^1(\nu)$ for the function ψ appearing in (iii) of theorem 3.4.1

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} for the 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:*

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

for an arbitrary $x_0 \in \mathcal{X}$.⁷

For $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ the Wasserstein distance of order p between μ and ν is defined as

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

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", also exist. 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".

Example 3.6.3. The Wasserstein distances induce some intuitive properties from a human point of view, such as when it comes to *Wasserstein barycenters*. A Wasserstein barycenter of measures $\{\mu_1, \dots, \mu_n\} \subset \mathcal{P}_p(\mathcal{X})$ is any measure $\mu \in \mathcal{P}_p(\mathcal{X})$ such that

$$\mu = \arg \min_{\mu' \in \mathcal{P}_p(\mathcal{X})} \sum_{i=1}^n W_p^p(\mu', \mu_i).$$

⁷Note that the space does not depend on the choice of x_0 .

Now assume you have two images of the digit 1, one on the left hand side of the image, the other on the right, and you are looking for an 'average 1' between the two. If you were to simply average over all pixels, you would get two transparent ones on both sides of the image – not a good solution. However, in the case $p > 1$, the unique Wasserstein barycenter would be a 1 in the middle of the image, which is consistent with what one would intuitively consider to be the average 1 (cmp. figure 3). Note that for the case $p = 1$, the Wasserstein barycenter would not be unique anymore, and both the two transparent ones as well as the 1 in the middle would indeed be Wasserstein barycenters.

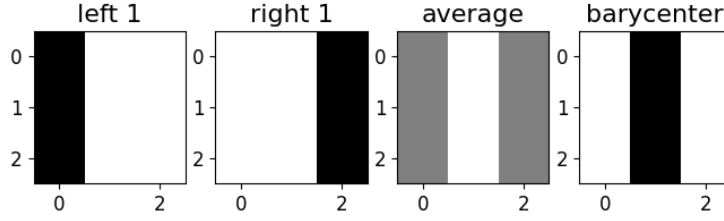


Figure 3: Two images of a 1 (left and second from left), their pixel average (second from right) and their unique Wasserstein barycenter for $p > 1$.

Remark 3.6.4 (Special Case W_1). In the case $p = 1$ the Wasserstein distance takes another special form. We know from theorem 3.5.1 that

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

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 $W_1(\mu, \nu)$:

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

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 3.6.5. 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 3.6.6 (Gluing). *Let (\mathcal{X}_i, μ_i) , $i \in \llbracket 3 \rrbracket$, 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 $i \in \llbracket 3 \rrbracket$.*

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 $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be the usual projection. By the disintegration theorem A.17 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.

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

i.e. the measures γ_x are concentrated on the fibres $\{x\} \times \mathcal{Y}$ a.e., meaning we can from now on consider γ_x to be a measure on \mathcal{Y} , and s.t. for any $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty]$ we have

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

We will write this as

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

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

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

Set

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

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

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

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

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

and as this holds in particular for test functions $\phi \in C_b(\mathcal{X}_1 \times \mathcal{X}_2)$, lemma A.19 implies that γ admits γ_{12} as its marginal on $\mathcal{X}_1 \times \mathcal{X}_2$. Similarly, one shows that the marginal of γ on $\mathcal{X}_2 \times \mathcal{X}_3$ is γ_{23} , which finishes the proof. \square

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

Theorem 3.6.7 (Wasserstein Distance is a Metric). *Let (\mathcal{X}, d) be a Polish space and $p \in [1, \infty)$. Then W_p defines a metric on $\mathcal{P}_p(\mathcal{X})$.*

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

- (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.12):

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

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 3.2.9). Then

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

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

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

which gives us $\mu_1 = \mu_2$ (cmp. lemma A.19). 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 3.2.9). Let $\mathcal{X}_i := \mathcal{X}$ for $i \in \llbracket 3 \rrbracket$ and define γ as in the gluing lemma 3.6.6, i.e.

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

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

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

where in the fourth step we used Minkowski's inequality (cmp. proposition A.20). \square

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 [41], theorem 6.18.

Theorem 3.6.8. *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 on finite, discrete spaces. 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: Say you are given an image $Q \in \mathbb{R}_{\geq 0}^{m \times n}$, where q_{ij} corresponds to the value of a single pixel, i.e. the higher q_{ij} , the more color the pixel contains. Then by setting $\mathcal{X} := \{x_1, \dots, x_{mn}\}$, we can define a measure μ corresponding to Q by setting

$$\mu = \sum_{k=1}^{mn} \frac{\text{vec}(Q)_k}{\sum_{i,j} q_{ij}} \delta_{x_k},$$

where oftentimes $\text{vec}(Q)$ will actually be considered to be the vector of concatenated rows of Q , instead of the columns, in practice.

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 [33].

We will start off with formulating the OT problem in the discrete setting. Let $\mathcal{X} = \{x_1, \dots, x_m\}$ and $\mathcal{Y} = \{y_1, \dots, y_n\}$ for some $n, m \in \mathbb{N}_{>0}$. Both spaces are equipped with the discrete topologies $\mathcal{T}(\mathcal{X}) = 2^{\mathcal{X}}$ and $\mathcal{T}(\mathcal{Y}) = 2^{\mathcal{Y}}$ resp., i.e. the Borel σ -algebras are equal to these power sets of \mathcal{X} and \mathcal{Y} : $\mathcal{B}(\mathcal{X}) = \mathcal{T}(\mathcal{X})$, $\mathcal{B}(\mathcal{Y}) = \mathcal{T}(\mathcal{Y})$. Further let $\mu \in \Delta^{m-1}$ and $\nu \in \Delta^{n-1}$ in the probability simplex. We will slightly abuse notation in the following, interchangeably considering μ and ν to be these vectors as well as the measures associated with them, namely $\sum_i \mu_i \delta_{x_i}$, $\sum_j \nu_j \delta_{y_j}$. This will make the notation a lot more readable. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a cost function and set $c_{ij} := c(x_i, y_j)$ for $i \in \llbracket m \rrbracket$, $j \in \llbracket n \rrbracket$. Then the Kantorovich problem 3.2.1 becomes:

Problem 3.7.1 (Discrete Optimal Transport Problem).

$$\begin{aligned} & \min_{\gamma \in \mathbb{R}_{\geq 0}^{m \times n}} \langle c, \gamma \rangle \\ \text{s.t. } & \gamma 1_n = \mu \\ & \gamma^\top 1_m = \nu \end{aligned}$$

Similarly, the dual problem 3.3.1 reads:

Problem 3.7.2 (Discrete Dual Optimal Transport Problem).

$$\begin{aligned} & \max_{f \in \mathbb{R}^m, g \in \mathbb{R}^n} \langle f, \mu \rangle + \langle g, \nu \rangle \\ \text{s. t. } & f + g \leq c \end{aligned}$$

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

$$\Pi(\mu, \nu) = \left\{ \gamma \in \mathbb{R}_{\geq 0}^{m \times n} : \sum_j \gamma_{ij} = \mu_i \text{ for all } i \in \llbracket m \rrbracket, \sum_i \gamma_{ij} = \nu_j \text{ for all } j \in \llbracket n \rrbracket \right\}.$$

This set is a polytope, as can more easily be seen from the following reformulation of the problem. In order to rewrite problem 3.7.1 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)

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

Then we can reformulate problem 3.7.1 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 [7], p. 4). We can also easily verify that the discrete dual problem 3.7.2 turns into the dual linear program. If we let $\psi \in \mathbb{R}^m$ and $\phi \in \mathbb{R}^n$ and set $y = \begin{bmatrix} \psi^\top & \phi^\top \end{bmatrix}^\top$, 3.7.2 becomes:

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

This is exactly the dual linear program to 9 (cmp. [7], 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.⁸ This can e.g. be found as Theorem 2.7. in [7], p. 65. Using this property, we can prove the following proposition.

Proposition 3.7.3. *Let $P \in \Pi(\mu, \nu)$ be a vertex of the polytope of feasible measures to problem 3.7.1. Then P does not have more than $m + n - 1$ nonzero entries. In particular, there exists an optimal solution to problem 3.7.1 with at most $m + n - 1$ nonzero entries.*

Proof. Define $V = \{1, \dots, m\}$, $V' = \{1', \dots, n'\}$ as two sets of nodes⁹ and let $G = (V \cup V', E)$ be

⁸Under some additional assumptions. These assumptions can for example be: There exists an optimal solution, and the polytope has at least one vertex. Both of these are fulfilled in our case.

⁹The primes in V' are simply meant to be able to disambiguate its elements from elements in V .

a directed graph with $E := \{(i, j') : i \in \llbracket m \rrbracket, j' \in \llbracket n \rrbracket\}$. Let P be a vertex of $\Pi(\mu, \nu)$. Then P corresponds to a "flow" in G , where we associate P_{ij} with the flow on $(i, j') \in E$. Let $E_P \subset E$ be the edges where $P > 0$. Now, if we can show that $(V \cup V', E_P)$ contains no undirected cycles, this proves that P has at most $m + n - 1$ nonzero entries as any graph with k nodes and no cycles can contain at most $k - 1$ edges.

Assume for the sake of contradiction that P does contain a cycle, i.e. there exists some $k \in \mathbb{N}$, $k \geq 2$, and indices $i_l \in \llbracket m \rrbracket$, $j_l \in \llbracket n \rrbracket$ for $l \in \llbracket k - 1 \rrbracket$, such that

$$H := \{(i_1, j'_1), (i_2, j'_1), (i_2, j'_2), \dots, (i_k, j'_k), (i_1, j'_k)\} \subset E_P.$$

Consider the directed cycle

$$\bar{H} := \{(i_1, j'_1), (j'_1, i_2), (i_2, j'_2), \dots, (i_k, j'_k), (j'_k, i_1)\}.$$

Let $\varepsilon > 0$ such that $\varepsilon < \min_{(i, j') \in E_P} P_{ij}$. Let $\mathcal{E} \in \mathbb{R}^{m \times n}$ be defined by $\mathcal{E}_{ij} = \varepsilon$ if $(i, j') \in \bar{H}$, $\mathcal{E}_{ij} = -\varepsilon$ if $(j', i) \in \bar{H}$ and $\mathcal{E}_{ij} = 0$ otherwise. Note that $\mathcal{E}1_n = 0$ and $\mathcal{E}^\top 1_m = 0$ by construction. Now define $Q = P + \mathcal{E}$ and $R = P - \mathcal{E}$, see figure 4.

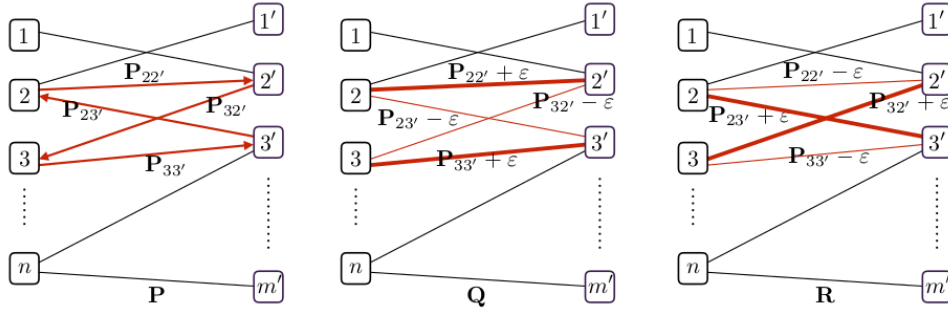


Figure 4: Constructing two transport plans Q and R from P .¹⁰

Then by construction, $Q, R \geq 0$ (entry-wise) and the marginals of Q and R equal those of P , hence $Q, R \in \Pi(\mu, \nu)$. This means $P = \frac{Q+R}{2}$ with $Q \neq P \neq R$, which contradicts P being a vertex.

Finally, note that as mentioned earlier, the optimum of problem 3.7.1 will occur at a vertex of $\Pi(\mu, \nu)$ (cmp. Theorem 2.7. in [7], p. 65), hence there is an optimal solution with at most $m + n - 1$ nonzero entries. \square

Leveraging this property, there exist multiple algorithms which can solve problem 3.7.1 precisely. For example, the *Network Simplex Algorithm* is a version of the well-known *Simplex Algorithm* which works particularly well in this case, cmp. also [33], 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 [31] was able to prove a polynomial complexity bound on the algorithm which was shortly after improved by Tarjan [39] 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* [27] or the *Auction Algorithm* [6], but none of them tends to be fast in high-dimensional spaces. This is why Cuturi proposed a different approach in his seminal work [11]: The so-called *Sinkhorn Algorithm*. We will see how it works in the following chapter.

¹⁰Source: [33], figure 3.1.

4 Sinkhorn Algorithm

In this section, we get to know the Sinkhorn algorithm, an iterative algorithm which computes an approximation of the transport plan and cost of the discrete optimal transport problem 3.7.1. We will discuss its advantages and disadvantages and pay particular attention to its initialization. Again, let $\mathcal{X} = \{x_1, \dots, x_m\}$ and $\mathcal{Y} = \{y_1, \dots, y_n\}$ for some $n, m \in \mathbb{N}_{>0}$ be two discrete spaces equipped with the discrete topologies $\mathcal{T}(\mathcal{X}) = 2^{\mathcal{X}}$ and $\mathcal{T}(\mathcal{Y}) = 2^{\mathcal{Y}}$ resp. Furthermore, let $\mu \in \Delta^{m-1}$ and $\nu \in \Delta^{n-1}$ in the probability simplex, and as in section 3.7 we consider μ and ν to be the measures $\sum_i \mu_i \delta_{x_i}$, $\sum_j \nu_j \delta_{y_j}$ when needed. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a cost function and set $c_{ij} := c(x_i, y_j)$ for $i \in \llbracket m \rrbracket$, $j \in \llbracket n \rrbracket$. The idea underlying the Sinkhorn algorithm is to introduce an *entropic regularizer* to the optimal transport problem. This will alter the problem and its solution, which means we will not be finding solutions to our regular optimal transport problem 3.7.1 anymore, but merely approximations. This drawback can be justified by the fact that the entropic problem comes with many useful properties the regular problem does not have.

4.1 Entropic Optimal Transport

Definition 4.1.1 (Entropy). *For a matrix $P = [p_{ij}]_{ij} \in \mathbb{R}^{m \times n}$, 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 negative. For entries $p_{ij} = 0$, we use the convention $0 \log 0 = 0$, as $x \log x \xrightarrow{x \rightarrow 0} 0$.

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. Our definition is similar to the definition in [33].¹¹ Other definitions exist, such as $H(P) = - \sum_{i,j} p_{ij} \log p_{ij}$, which can e.g. be found in [11]. 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 or \ln .

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.

¹¹In [33], $H(P) := -\infty$ also if an entry is merely equal to 0. However, it seems like this might be a typo, as with this definition of entropy, some of the results in [33] would not hold.

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

Problem 4.1.4 (Entropic Optimal Transport Problem). For $\varepsilon > 0$, the entropic optimal transport problem is defined as:

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

The term $-\varepsilon H(\gamma)$ is referred to as the entropic regularizer, and ε as the regularizing constant.

Remark 4.1.5. As the entropy is a 1-strongly convex function¹² on all transport plans (since $\partial^2 H(P) = -\text{diag}(\frac{1}{p_{ij}})$ 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 $\varepsilon \rightarrow \infty$, the unique solution to (10) converges to the transport plan of maximum entropy, which, as we just saw, is the trivial coupling, and as $\varepsilon \rightarrow 0$, the solution to (10) indeed converges to the maximum entropy optimal coupling of the unregularized problem.¹³

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

$$\gamma_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} \arg \min \left\{ -H(\gamma) : \gamma \in \Pi(\mu, \nu), \langle c, \gamma \rangle = \min_{\gamma' \in \Pi(\mu, \nu)} \langle c, \gamma' \rangle \right\} \quad (11)$$

and

$$\gamma_\varepsilon \xrightarrow{\varepsilon \rightarrow \infty} \mu \nu^\top.$$

Proof. This proof follows that of proposition 4.1 in [33]; note, however, that for this proof to work, we need to have H defined as in definition 4.1.1 and not as in [33]. Consider a sequence $(\varepsilon_l)_{l \in \mathbb{N}}$ in $\mathbb{R}_{>0}$ s.t. $\varepsilon_l \xrightarrow{l \rightarrow \infty} 0$. Denote by γ_l the unique solution to problem 4.1.4 for $\varepsilon = \varepsilon_l$. Since $\Pi(\mu, \nu)$ is bounded, we can extract a subsequence (which we will also denote by $(\gamma_l)_l$ for simplicity) which converges to some $\gamma^* \in \mathbb{R}^{m \times n}$, and since $\Pi(\mu, \nu)$ is closed we have $\gamma^* \in \Pi(\mu, \nu)$. Now let $\gamma \in \Pi(\mu, \nu)$ be optimal for the unregularized primal problem 3.7.1. By optimality of γ for 3.7.1 and optimality of γ_l for 4.1.4, we have

$$0 \leq \langle c, \gamma_l \rangle - \langle c, \gamma \rangle \leq \varepsilon_l (H(\gamma_l) - H(\gamma)). \quad (12)$$

Since H is continuous, we know that $H(\gamma_l) - H(\gamma)$ is bounded, and taking the limit $l \rightarrow \infty$ in (12) shows that $\langle c, \gamma^* \rangle = \langle c, \gamma \rangle$, i.e. γ^* is feasible for (11). Dividing by ε_l in (12) and taking the limit again shows that $H(\gamma) \leq H(\gamma^*)$, which again follows from continuity of H . Also, since the solution to problem 4.1.4 is unique for any ε by strict convexity of H , and since H is continuous, the entire sequence γ_l has to converge to γ^* (and not just a subsequence). For the limit $\varepsilon \rightarrow \infty$, a similar proof shows that one can consider the problem

$$\min_{\gamma \in \Pi(\mu, \nu)} -\varepsilon H(\gamma)$$

instead, which is solved by $\mu \nu^\top$. □

¹²A 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 .

¹³We 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.

Remark 4.1.7. With a different definition to the entropic optimal transport problem, Cuturi [11] showed that, similar to the Wasserstein distances, the optimal value of the entropic problem defines a metric on the distributions (cmp. [11], 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*

$$K_{ij} = \exp\left(-\frac{c_{ij}}{\varepsilon}\right), \quad i \in \llbracket m \rrbracket, j \in \llbracket n \rrbracket.$$

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*

$$\gamma_{ij} = u_i K_{ij} v_j, \quad i \in \llbracket m \rrbracket, j \in \llbracket n \rrbracket, \quad (13)$$

where K is the Gibbs kernel and $u \in \mathbb{R}_{>0}^m$ and $v \in \mathbb{R}_{>0}^n$ 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 idea of the proof is to use Lagrange multipliers to derive equation (13) (cmp. proposition A.21). As the entropy of γ in problem 4.1.4 becomes $-\infty$ once some element in γ is negative, we can drop the constraint that $\gamma \geq 0$ and are left with the constraints $\gamma 1_n = a$ and $1_m^\top \gamma = b^\top$. This means we have $m + n$ equality constraints. Writing the equality constraints as functions $h_i : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ for $i \in \llbracket m + n \rrbracket$,

$$h_i(\gamma) := \sum_{j=1}^n \gamma_{ij} - a_i, \quad i \in \llbracket m \rrbracket, \quad h_{m+j}(\gamma) := \sum_{i=1}^m \gamma_{ij} - b_j, \quad j \in \llbracket n \rrbracket,$$

we can see that the gradients $\nabla h_i(\gamma)$ for $i \in \llbracket m + n - 1 \rrbracket$ are linearly independent, regardless of whether γ is optimal or not, as

$$(\nabla h_k(\gamma))_{ij} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases}, \quad k \in \llbracket m \rrbracket, \quad (\nabla h_{m+k}(\gamma))_{ij} = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}, \quad k \in \llbracket n \rrbracket, \quad (14)$$

i.e. they are equal to the $m + n$ $m \times n$ -dimensional matrices where exactly one row resp. column is equal to 1, and all other entries are equal to 0. Note that adding $\nabla h_{m+n}(\gamma)$ would indeed make the gradients linearly dependent, however. Now we can write the objective in (10) in terms of a function $F : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ via

$$F(\gamma) := \begin{cases} \langle \gamma, C + \varepsilon \log \gamma - \varepsilon \rangle, & \gamma_{ij} \geq 0 \text{ for all } i \in \llbracket m \rrbracket, j \in \llbracket n \rrbracket, \\ \infty, & \text{otherwise,} \end{cases}$$

again with the convention $0 \log 0 = 0$. We can assume without loss of generality that we are only dealing with matrices $\gamma_{ij} \geq 0$ for all $i \in \llbracket m \rrbracket, j \in \llbracket n \rrbracket$, as mentioned earlier. Now assume that γ is optimal for (10). By proposition A.21, we know there exist unique Lagrange multipliers $f \in \mathbb{R}^m$ and

$\tilde{g} \in \mathbb{R}^{n-1}$ such that¹⁴

$$\nabla F(\gamma) + \sum_{i=1}^m f_i \nabla h_i(\gamma) + \sum_{j=1}^{n-1} g_j \nabla h_{m+j}(\gamma) = 0.$$

Computing

$$\nabla F(\gamma) = \nabla_{\gamma} \langle \gamma, C + \varepsilon \log \gamma - \varepsilon \rangle C + \varepsilon (\log \gamma + 1) - \varepsilon = C + \varepsilon \log \gamma,$$

and using (14) we get

$$C_{ij} + \varepsilon \log \gamma_{ij} + f_i + g_j = 0 \text{ for all } i \in \llbracket m \rrbracket, j \in \llbracket n-1 \rrbracket, \quad (15)$$

$$C_{in} + \varepsilon \log \gamma_{in} + f_i = 0 \text{ for all } i \in \llbracket m \rrbracket, \quad (16)$$

which is equivalent to

$$\begin{aligned} \gamma_{ij} &= e^{f_i/\varepsilon} e^{-c_{ij}/\varepsilon} e^{g_j/\varepsilon} \text{ for all } i \in \llbracket m \rrbracket, j \in \llbracket n-1 \rrbracket, \\ \gamma_{ij} &= e^{f_i/\varepsilon} e^{-c_{ij}/\varepsilon} \text{ for all } i \in \llbracket m \rrbracket. \end{aligned}$$

Hence, setting $g_n = 0$, we get

$$\gamma = \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}),$$

and setting $u := f/\varepsilon$ and $v := g/\varepsilon$ gives us a representation as in (13). Also note that uniqueness of u and v up to a scaling constant immediately follows from the proof: We showed that for $g_n = 0$, i.e. $v_n = 1$, this representation is unique. If we chose a different value for g_n , say, $k \in \mathbb{R}$, then from (16) it follows that we have to deduct k from all f_i , $i \in \llbracket m \rrbracket$. Then, in turn, it follows from (15) that we have to add k to all g_j , $j \in \llbracket n-1 \rrbracket$. Hence, this again gives us a unique solution, where we added k to g and subtracted it from f , which corresponds to multiplying v by $e^{k/\varepsilon}$ and u by its inverse $e^{-k/\varepsilon}$. \square

As does the unconstrained problem, the entropic version also comes with its own dual problem. We will derive it using methods from nonlinear programming (cmp. [5], particularly chapter 5.1,¹⁵ and also [30]). The Lagrangian associated with problem 4.1.4, in the primal variable γ and the dual variables f and g , reads

$$\mathcal{L}(\gamma, f, g) = \langle c, \gamma \rangle + \varepsilon \langle \gamma, \log \gamma - 1 \rangle + \langle f, \mu - \gamma 1_n \rangle + \langle g, \nu - \gamma^\top 1_m \rangle.$$

This gives us

$$L^\varepsilon(\mu, \nu) = \inf_{\gamma} \sup_{f, g} \mathcal{L}(\gamma, f, g),$$

and the dual problem can be obtained by interchanging the infimum and supremum:

$$D^\varepsilon(\mu, \nu) = \sup_{f, g} \min_{\gamma} \sum_{ij} \gamma_{ij} (c_{ij} + \varepsilon \log \gamma_{ij} - \varepsilon - f_i - g_j) + \sum_i f_i \mu_i + \sum_j g_j \nu_j. \quad (17)$$

¹⁴We choose a different sign for the Lagrange multipliers than in proposition A.21 because this way, it will turn out that f and g are actually the dual potentials from the entropic dual problem 4.1.10, as we will soon see.

¹⁵Linear and nonlinear programs always come in various flavours. In this chapter, Bertsekas considers inequality constraints $g_j(x) \leq 0$. Since we are dealing with equality constraints in our case, one can simply expand the inequality constraints by further inequality constraints $-g_j(x) \leq 0$, which is then equivalent to our equality constraints. This means one can drop assumptions like the Lagrange multiplier vector being nonnegative – cmp. 5.1.1 in [5] – or the inequality constraints on the dual problem therein, see 5.1.2.

Since for a given pair (f, g) , we take the minimum over all transport plans γ , first-order conditions yield

$$\begin{aligned} c_{ij} - f_i - g_j + \varepsilon \log \gamma_{ij} &= 0, \\ \text{i.e. } \gamma_{ij} &= e^{f_i/\varepsilon} e^{-c_{ij}/\varepsilon} e^{g_j/\varepsilon} \text{ for all } i \in \llbracket m \rrbracket, j \in \llbracket n \rrbracket, \end{aligned}$$

similarly to what we have seen in the proof of proposition 4.1.9. Plugging this back into (17) yields the following dual problem:

Problem 4.1.10 (Entropic Dual Problem). The entropic dual problem is defined as:

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

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

$$L^\varepsilon(\mu, \nu) = D^\varepsilon(\mu, \nu). \quad (18)$$

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

$$(u, v) = (e^{f/\varepsilon}, e^{g/\varepsilon}). \quad (19)$$

Proof. We know that $L^\varepsilon(\mu, \nu) \geq D^\varepsilon(\mu, \nu)$ (cmp. proposition 5.1.3 in [5]). Hence, it suffices to show that there exist optimal γ for the primal and (f, g) for the dual such that there is equality in order to prove duality as in (18). Let γ be optimal for the primal problem 4.1.4. By proposition 4.1.9 and its proof we know that we then have for some $f \in \mathbb{R}^m$ and $g \in \mathbb{R}^n$:

$$\begin{aligned} L^\varepsilon(\mu, \nu) &= \left\langle c, \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}) \right\rangle + \varepsilon \left\langle \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}), \log \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}) - 1 \right\rangle \\ &= \left\langle c, \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}) \right\rangle + \left\langle \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}), f 1_n^\top + 1_m g^\top - c - \varepsilon \right\rangle \\ &= \left\langle \text{diag}(e^{f/\varepsilon}) K \text{diag}(e^{g/\varepsilon}), f 1_n^\top + 1_m g^\top - \varepsilon \right\rangle \\ &= \langle f, \mu \rangle + \langle g, \nu \rangle - \varepsilon \left\langle e^{f/\varepsilon}, K e^{g/\varepsilon} \right\rangle. \end{aligned}$$

This shows that (f, g) is an optimizer of the dual problem 4.1.10, and (19) holds as well, which follows from the representation of the optimal plan in (13) in proposition 4.1.9. \square

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 useful property that the unregularized problem did not have, and one we will make use of.

A solution (f, g) of the unregularized problem 3.7.2 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 3.7.2 and $(f^\varepsilon, g^\varepsilon)$ a solution to the regularized dual problem 4.1.10 for some $\varepsilon > 0$. Then $(f^\varepsilon, g^\varepsilon)$ is feasible for the unregularized problem, i.e. $f^\varepsilon + g^\varepsilon \leq c$, and*

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

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

$$\langle f^\varepsilon, \mu \rangle + \langle g^\varepsilon, \nu \rangle \xrightarrow{\varepsilon \rightarrow 0} \langle f, \mu \rangle + \langle g, \nu \rangle. \quad (20)$$

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

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

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

$$\langle f, \mu \rangle + \langle g, \nu \rangle \geq \langle f^\varepsilon, \mu \rangle + \langle g^\varepsilon, \nu \rangle.$$

This gives us

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

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 $(f^\varepsilon, g^\varepsilon)$. This also implies (20), as

$$\begin{aligned} 0 &\geq \langle f^\varepsilon, \mu \rangle + \langle g^\varepsilon, \nu \rangle - \langle f, \mu \rangle + \langle g, \nu \rangle \\ &\geq -\varepsilon \left[\langle e^{f/\varepsilon}, K e^{g/\varepsilon} \rangle - \langle e^{f^\varepsilon/\varepsilon}, K e^{g^\varepsilon/\varepsilon} \rangle \right] \geq -\varepsilon mn. \end{aligned}$$

□

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):

$$\begin{aligned} u \odot K v &= \mu, \\ v \odot K^\top u &= \nu. \end{aligned}$$

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 $v^0 \in \mathbb{R}_{>0}^n$ (which can be initialized arbitrarily or according to some initialization scheme), the updates of the algorithm are as follows:

$$u^{l+1} = \frac{\mu}{K v^l}, \quad v^{l+1} = \frac{\nu}{K^\top u^{l+1}}, \quad l = 0, 1, 2, \dots, \quad (21)$$

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

Remark 4.2.1. The iterations (21) first appeared long before Sinkhorn and Knopp [38] proved their convergence in 1967; Yule [43] mentioned them in 1912 already. They have been known under various names such as *iterative proportional fitting procedure (IPFP)* [14] and *RAS* [4] throughout the years, and were e.g. used to solve matching problems in economics [20], before they received a boost of attention following Cuturi's paper *Sinkhorn distances: lightspeed computation of optimal transport* [11].

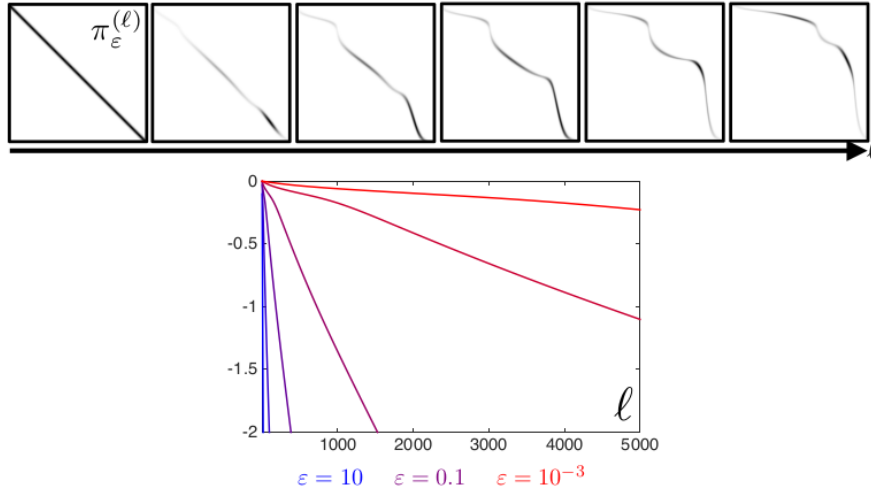


Figure 5: Top: Evolution of $\pi_\varepsilon^{(l)} := \text{diag}(u^l)K\text{diag}(v^l)$ as l increases, for $\varepsilon = 0.1$, $c(x, y) = |x - y|^2$ and one-dimensional distributions on $[0, 1]$. Bottom: Impact of the regularizing constant ε on the convergence rate of the algorithm, measured in terms of the marginal constraint violation on ν , $\log \left(\left\| 1_m^\top \pi_\varepsilon^{(l)} - \nu^\top \right\|_1 \right)$.¹⁶

¹⁶Source: [33], figure 4.5.

Proposition 4.2.2. *The iterates u^l and v^l as in (21) converge to the up to a constant uniquely defined u and v in proposition 4.1.9 as $l \rightarrow \infty$.*

Proof. This statement goes back to the original work of Sinkhorn and Knopp [38], where convergence is proven. Another proof using a Hilbert projective metric can be found in [18]. \square

Remark 4.2.3. Note that with varying initializations 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 $\lim_{l \rightarrow \infty} \text{diag}(u^l)K\text{diag}(v^l)$ will always be the unique solution from proposition 4.1.9.

This algorithm lies at the heart of the seminal work of Cuturi [11]. The algorithm from his paper can be seen in figure 6, while the equivalent algorithm using our iteration as above can be seen in algorithm 1.¹⁷

Algorithm 1 Sinkhorn Algorithm

```

1: in  $c \in \mathbb{R}^{m \times n}$ ,  $\varepsilon > 0$ ,  $\mu \in \Delta_{>0}^{m-1}$ ,  $\nu \in \Delta_{>0}^{n-1}$ 
2: initialize  $v^0$  (e.g.  $v^0 \leftarrow 1_n$ ),  $l \leftarrow 0$ ,  $K \leftarrow e^{-c/\varepsilon}$ 
3: repeat
4:    $u^{l+1} \leftarrow \mu ./ K v^l$ 
5:    $v^{l+1} \leftarrow \nu ./ K^\top u^{l+1}$ 
6:    $l \leftarrow l + 1$ 
7: until stopping criterion is met
8:  $\gamma \leftarrow \text{diag}(u^l)K\text{diag}(v^l)$ 
9: out  $\gamma$ ,  $\langle \gamma, c \rangle$ 

```

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

```

Input M,  $\lambda$ , r, c.
I=(r>0); r=r(I); M=M(I,:); K=exp(-lambda*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_M^lambda(r,c)=sum(u.*(K.*M)*v)

```

Figure 6: Sinkhorn algorithm as in [11].

Cuturi uses M for c , λ for $\frac{1}{\varepsilon}$ and r, c for μ, ν . 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 5.

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 c and regularizing

¹⁷The expression $./$ denotes entry-wise division, and $.*$ is entry-wise multiplication.

constant ε , and a collection of pairs of distributions $(\mu_i, \nu_i)_{i \in \llbracket k \rrbracket}$, writing $A := \begin{bmatrix} \mu_1 & \dots & \mu_k \end{bmatrix}$, $B := \begin{bmatrix} \nu_1 & \dots & \nu_k \end{bmatrix}$ we can solve these problems simultaneously by initializing some $V^0 \in \mathbb{R}_{>0}^{n \times k}$ and updating $U, V \in \mathbb{R}_{>0}^{n \times k}$ as follows:

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

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 patterns [42], [15]. Also, choosing the regularizing constant 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 (21). For similar reasons, Sinkhorn does not support computations on distributions that contain zeros or very small values. To some extent, these problems can be dealt with by shifting computations to the log domain. Details can be found in [33], 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 convergence is indeed guaranteed, initialization matters in terms of *convergence speed*. If for example v^0 is already very close to some optimal v , this initialization will lead to much faster convergence than a random one.¹⁸ Comparatively little attention has been paid to improving initialization of Sinkhorn's algorithm. Thornton and Cuturi [40] 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 which is based on the Wasserstein distance approximation that the dual potential approximation yields; we will see how exactly in section 5.5. This has the clear advantage that you can simply 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;

¹⁸While intuitively clear, this will also become empirically evident in section 6.

also see section 5.5. 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.

5 Sinkhorn-NN Hybrid Algorithm

We now present our Sinkhorn-NN hybrid algorithm. The main idea is: Train a neural network to predict the dual potential of the discrete unconstrained optimal transport problem 3.7.1 and then use that to initialize the Sinkhorn algorithm 1. First, we will have a more detailed look at the idea itself and why it works in section 5.1, before going through the implementation in detail, with particular attention being paid to training and testing data generation in sections 5.2 – 5.3, the network architecture in 5.4, and training in 5.6. Additionally, in section 5.5, we answer some important questions regarding the algorithm.

As before, let $\mathcal{X} = \{x_1, \dots, x_m\}$ and $\mathcal{Y} = \{y_1, \dots, y_n\}$ for some $n, m \in \mathbb{N}_{>0}$ be two discrete spaces equipped with the discrete topologies $\mathcal{T}(\mathcal{X}) = 2^{\mathcal{X}}$ and $\mathcal{T}(\mathcal{Y}) = 2^{\mathcal{Y}}$ resp. Furthermore, let $\mu \in \Delta^{m-1}$ and $\nu \in \Delta^{n-1}$ in the probability simplex, and as usual we will abuse notation by sometimes referring to the measures $\sum_i \mu_i \delta_{x_i}$, $\sum_j \nu_j \delta_{y_j}$ by μ and ν as well. Let $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a cost function and set $c_{ij} := c(x_i, y_j)$ for $i \in \llbracket m \rrbracket$, $j \in \llbracket n \rrbracket$. Additionally, we define $K := e^{-c/\varepsilon}$ for some regularizing constant $\varepsilon > 0$; remember this is to be understood in an element-wise fashion, cmp. section 2.

5.1 A Trained Initialization for the Sinkhorn Algorithm

Ultimately, we want to be able to quickly approximate optima of 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 3.7.1 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 primal problem can be recovered via $\gamma = \text{diag}(u)K\text{diag}(v)$, see proposition 4.2.2. 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 3.7.2 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 $v = e^{g^\varepsilon/\varepsilon}$ for a solution $(f^\varepsilon, g^\varepsilon)$ of the entropic dual problem 4.1.10, cmp. proposition 4.1.11. Hence, we will let a neural network learn to approximate the mapping $(\mu, \nu) \mapsto (f, g)$ which maps two distributions to a solution of the unregularized dual problem.¹⁹ In fact, it suffices to consider the mapping $p : (\mu, \nu) \mapsto f$ as we know that at optimality, we can recover g from f via $g = f^c$, cmp. theorem 3.5.1.²⁰ We will do so using training data containing ground truth dual potentials. After training is finished, given two

¹⁹Note this is not a function, as there can exist multiple optima for the dual. However, as we will see soon, we employ multiple constrictions to reduce the degree of freedom this mapping has for a given input.

²⁰We 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.

distributions $\mu \in \Delta_{>0}^{m-1}$ and $\nu \in \Delta_{>0}^{n-1}$, we can compute $f \approx \text{net}(\mu, \nu)$, $g = f^c$, $v^0 = e^{g/\varepsilon}$ for a given $\varepsilon > 0$, and use this vector v^0 as a starting vector for the Sinkhorn algorithm. However, if g contains large or small values, this will quickly lead to entries in v^0 being 0 or ∞ due to numerical rounding errors. As we have seen, the Sinkhorn algorithm only works if v is positive everywhere, and obviously entries being ∞ needs to be prevented as well. Hence, we will bound each entry in v^0 from below by a small constant b_1 and from above by a large constant b_2 . A pseudocode of the Sinkhorn-NN hybrid algorithm can be seen in algorithm 2.

Algorithm 2 Sinkhorn-NN Hybrid Algorithm

```

1: in  $c \in \mathbb{R}^{m \times n}$ 
2: generate training data  $d = (\mu_{\text{train}}, \nu_{\text{train}}, f_{\text{train}}) \in \Delta^{m-1} \times \Delta^{n-1} \times \mathbb{R}^m$ 
3: initialize  $\text{net}_\theta$  with  $(m+n)$ -dim. input and  $m$ -dim. output with parameters  $\theta$ 
4: train  $\text{net}_\theta$  on  $d$  with  $\text{loss}(\mu_{\text{train}}, \nu_{\text{train}}) \leftarrow \text{MSE}(\text{net}_\theta(\mu_{\text{train}}, \nu_{\text{train}}), f_{\text{train}})$ 
5: in  $\varepsilon > 0$ ,  $\mu \in \Delta_{>0}^{m-1}$ ,  $\nu \in \Delta_{>0}^{n-1}$ ,  $b_1 \in \mathbb{R}$ ,  $b_2 \in \mathbb{R}$ 
6:  $g \leftarrow c\text{-transform}(\text{net}_\theta(\mu, \nu))$ 
7:  $v^0 \leftarrow e^{g/\varepsilon}$ ,  $l \leftarrow 0$ ,  $K \leftarrow e^{-c/\varepsilon}$ 
8:  $v^0 \leftarrow \max\{b_1, \min\{b_2, v^0\}\}$ 
9: repeat
10:    $u^{l+1} \leftarrow \mu ./ K v^l$ 
11:    $v^{l+1} \leftarrow \nu ./ K^\top u^{l+1}$ 
12:    $l \leftarrow l + 1$ 
13: until stopping criterion is met
14:  $\gamma \leftarrow \text{diag}(u^l) K \text{diag}(v^l)$ 
15: out  $\gamma$ ,  $\langle c, \gamma \rangle$ 

```

Note that once the training in steps 1 – 4 has completed, steps 5 – 15 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. Furthermore, our algorithm keeps two of the main advantages of the regular Sinkhorn algorithm that we have discussed in section 4.2: As neural networks can be parallelized in a similar fashion (i.e. computing outputs for multiple inputs at once via matrix multiplications), algorithm 2 allows for the same parallelization as the Sinkhorn algorithm 1. Also, our outputs are still differentiable in the inputs, as one can differentiate through the neural network as well.²¹

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/SinkhornNNHybrid>.

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 $p : \Delta^{m-1} \times \Delta^{n-1} \rightarrow \mathbb{R}^m$, $(\mu, \nu) \mapsto f$, and not only train the network on a small subset of $\Delta^{m-1} \times \Delta^{n-1}$. Thus, we cannot train

²¹To be precise, the maximum and minimum functions in step 8 are only differentiable almost everywhere, but that suffices in practice. They are implemented using the `torch.where` function, which also supports differentiation.

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, but does not work particularly 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 $\mu_i = 0$ for some i , then changing f_i does not alter the value of $\langle f, \mu \rangle$. Thus, enforcing all data points to be larger than some small threshold larger than 0 reduces the degree of freedom the mapping p has for a given input, and is not restrictive of the problem as the Sinkhorn algorithm only works on strictly positive data anyways.²² Second, by the law of large numbers (cmp., e.g. [9], theorem 10.10.22), we know that the average over any patch of the distribution will converge to 1 over the dimension of the distribution as the size of the patch grows. This means that particularly in larger dimensions, the mass of the distribution will become ‘evenly spread’, which results in the transport distances being almost identical between all distributions; a property which is not desirable for our training data, as the network needs distinguishable samples in order to efficiently learn from them. 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 makes the distribution’s mass concentrate on fewer data points. It ensures that the transport distances between samples are sufficiently large, and the samples differ more distinctly from one another. See figure 7 for a visualization of the effect of k .

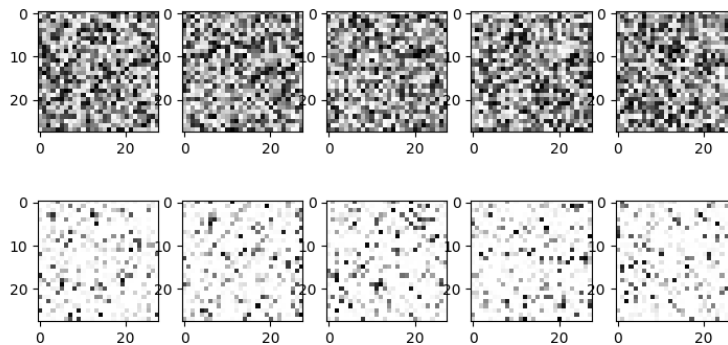


Figure 7: Top: data with $k_1 = k_2 = 0$. Bottom: our training data with $k_1 = 3$ and $k_2 = 0.001$.

Another important factor in producing data is the following: As we have seen, the dual value $\langle f, \mu \rangle + \langle f^c, \nu \rangle$ 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. In practice, we used the `emd` function in the POT package.

Now, we will have a look at how precisely we implemented this idea. We will be dealing with 28×28 -

²²As a random number between 0 and 1 is almost surely greater than 0 anyways, and hence 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.

Algorithm 3 Training Data Generation

```

1: in  $k_1 > 1, k_2 > 0$ 
2:  $\text{data} \leftarrow \text{list}()$ 
3: for  $i = 1, 2, \dots, N$  do
4:    $\mu \in [0, 1]^m, \nu \in [0, 1]^n$  random
5:    $\mu \leftarrow \mu^{k_1}, \nu \leftarrow \nu^{k_1}$ 
6:    $\mu \leftarrow \mu + k_2, \nu \leftarrow \nu + k_2$ 
7:    $\mu \leftarrow \frac{\mu}{\sum_i \mu_i}, \nu \leftarrow \frac{\nu}{\sum_i \nu_i}$ 
8:    $f \leftarrow \text{DualPotential}(\mu, \nu)$ 
9:    $f \leftarrow f - \frac{\sum_i f_i}{m}$ 
10:   $\text{append}(\text{data}, (\mu, \nu, f))$ 
11: end for
12: out  $\text{data}$ 

```

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.²³

The cost matrix used throughout all experiments is the squared Euclidean distance, i.e. the cost to get from μ_{ij} (considering μ to be 28×28 -dimensional instead of 784-dimensional) to $\nu_{i'j'}$ 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,²⁴ but any other cost function could have also been used.²⁵

5.3 Test Data

As we want to be able to generalize to any dataset, we will test on four different test datasets which were chosen to be varying in appearance and structure, each containing 10.000 samples. One is generated in the same way as the training data, one contains drawings of teddy bears from the Quick, Draw! dataset, one equals the test dataset of MNIST (hand-written digits), and the last one is a black and white version of the CIFAR10 test dataset (which contains images of boats, cars etc.). For each of them, 10.000 distributions have been generated, and for each dataset sample two of these have been chosen at random. For the same reasons as with the training data, all test datasets were modified with

²³Note that these values are very specific to the sample size. With 14×14 -dimensional data, for instance, $k_1 = 2$ proved to be better.

²⁴Another common choice would be the regular Euclidean distance, yielding the Wasserstein-1 distance. As we have seen in remark 3.6.4, 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 $\mu - \nu$ into the network instead of (μ, ν) , resulting in an input size of 784 instead of 1568.

²⁵However, 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.

a constant $k_2 = 0.001$ as in algorithm 3.²⁶ In addition to the two distributions and the dual potential, the test datasets also contain the transport cost between the two distributions. We again used the `emd` function in the POT package for computing the dual potentials and transport costs. We will oftentimes abbreviate them by 'random', 'teddies', 'MNIST' and 'CIFAR'. Figure 8 shows samples from each of the test datasets.

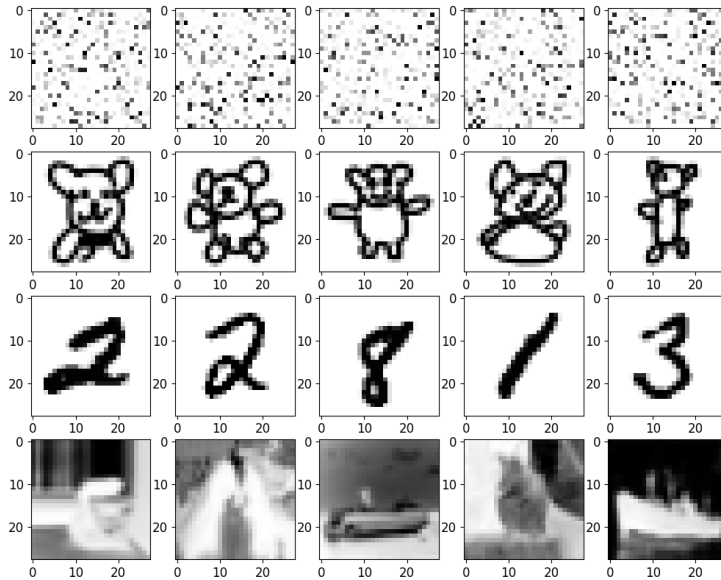


Figure 8: Test datasets 'random', 'teddies', 'MNIST' and 'CIFAR' (from top to bottom).

5.4 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.²⁷ The first two layers contain a batch normalization layer (see [22]) and a ReLU activation, 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 [25], and varying learning rates depending on the experiment.

²⁶Adding this much mass to a 28×28 -dimension probability distribution would greatly distort the original distribution, as it corresponds to adding a total of 0.784 to the distribution, i.e. almost doubling its mass. However, we do not have to worry about this problem, as the Quick, Draw!, MNIST and CIFAR10 datasets all come with an average mass of 40.000, 100 and 1500 per distribution resp., i.e. they are not normalized yet.

²⁷Note that all training was done on this CPU, as no GPU was available. If shifted to GPU, all training will be a lot faster than the values reported here.

Unless stated otherwise, in experiments with 100,000 training samples the learning rate will be set to 0.005, and in experiments with 1 million training samples it will decay from 0.005 to 0.0005. For comparison, the default learning rate for the Adam optimizer is 0.001. Since we know that our optimal transport distances are equal to the squared Wasserstein-2 distance which is a metric (cmp. theorem 3.6.7), we know it is symmetric. Symmetry is easy to enforce: Instead of returning $\text{net}(\mu, \nu)$, we can return $\frac{\text{net}(\mu, \nu) + \text{net}(\nu, \mu)^c}{2}$, i.e. computing the network's output for (μ, ν) and then switching the order of the input distributions. Note that we have to compute the c -transform of $\text{net}(\nu, \mu)$ as if we switch the distributions' order, this means instead of maximizing $\langle \cdot, \mu \rangle + \langle \cdot, \nu \rangle$ we now maximize $\langle \cdot, \nu \rangle + \langle \cdot, \mu \rangle$, and if (f, g) maximizes the former, we have $g = f^c$ by theorem 3.5.1 and the maximum equals $\langle f, \mu \rangle + \langle f^c, \nu \rangle$, with f appearing in the first scalar product. However, this means (g, f) is optimal for the latter, again with $g = f^c$, and $\langle f^c, \nu \rangle + \langle f, \mu \rangle$ 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%.

Furthermore, we know that the network's output should be a c -concave function. Again, this is easy to enforce: Instead of returning $\text{net}(\mu, \nu)$, we can return its ' c -concavification' $\text{net}(\mu, \nu)^{cc}$ 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

$$W_2^2(\mu, \nu) \approx \langle \text{net}(\mu, \nu), \mu \rangle + \langle \text{net}(\mu, \nu)^c, \nu \rangle.$$

The MSE 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.²⁸ 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.

5.5 Why Not...?

By now, there are a few important questions one might be wondering which we have to address, such as: Why don't we use the Sinkhorn algorithm in creating data, as this would give us the potential from

²⁸At 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 5.5.

the regularized problem, 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? 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? On the one hand, we know by proposition 4.1.13 that the solution of the unregularized dual problem 3.7.2 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 regularizing constant $\varepsilon > 0$ for generating the data. This might increase convergence speed for *that particular* regularizing constant. But we want a universal network which can be used with the Sinkhorn algorithm with varying regularizing constants.

One advantage of using the Sinkhorn algorithm would be that the solution of the entropic dual problem is unique up to a constant – unlike the solution to the unregularized problem. This means that the mapping which maps a pair of distributions to the dual potential can be turned into a function, if we ensure uniqueness w.r.t. said constant, which can, e.g., be achieved by only considering potentials f in the entropic dual problem 4.1.10 which sum to 0. However, while the Sinkhorn algorithm does converge to the solution in theory, in practice we would have to stop after a finite number of iterations, losing this uniqueness property.

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

$$\min_{\gamma \in \Pi(\mu, \nu)} = \langle f, \mu \rangle + \langle f^c, \nu \rangle.$$

This means we could approximate the transport distance between μ and ν by calculating $\langle \text{net}(\mu, \nu), \mu \rangle + \langle \text{net}(\mu, \nu)^c, \nu \rangle$. So why do we need to feed $\text{net}(\mu, \nu)$ 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 vectors 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 mapping $p : (\mu, \nu) \mapsto f$ 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 as can be seen in table 1.

²⁹Network trained for one million training samples, which saturates learning as we will see in section 5.6. Tested on 20 sets with 50 samples each. Note how for very small numbers of iterations, the network without Sinkhorn is actually superior. For comparison: The respective values for the Sinkhorn algorithm with default initialization for 1, 50, 200, 1000, and 2400 iterations are 5.322 ± 0.083 , 3.098 ± 0.079 , 2.098 ± 0.079 , 0.744 ± 0.043 , and 0.145 ± 0.008 , showing that the improved approximations compared to the network by itself are not due to the Sinkhorn algorithm alone, but also stem from the initialization. However, we will have a closer look at this in the following section 6.

network	Sinkhorn with network initialization				
	1 iteration	50 iterations	200 iterations	1000 iterations	2400 iterations
1.706 ± 0.021	3.654 ± 0.087	2.038 ± 0.083	1.338 ± 0.075	0.425 ± 0.036	0.079 ± 0.007

Table 1: L^1 errors on Wasserstein distance with 95% confidence intervals on 'random' test data for the network itself, compared to network+fixed number of Sinkhorn iterations.²⁹

Why not use a loss on the transport distance? For a sample (μ, ν, f) from the training data, the loss we use is $\text{loss}(\mu, \nu) = \text{MSE}(\text{net}(\mu, \nu), f)$. However, since we know that the expression $\langle \cdot, \mu \rangle + \langle \cdot, \nu \rangle$ 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

$$\text{loss}(\mu, \nu) = -(\langle \text{net}(\mu, \nu), \mu \rangle + \langle \text{net}(\mu, \nu)^c, \nu \rangle) \quad (22)$$

as a loss. This is in fact the loss used by Amos et al. [2]. It 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 optimal transport problems for generating training data. However, as can be seen in figure 9, this leads to significantly worse approximations on the potential than our approach.

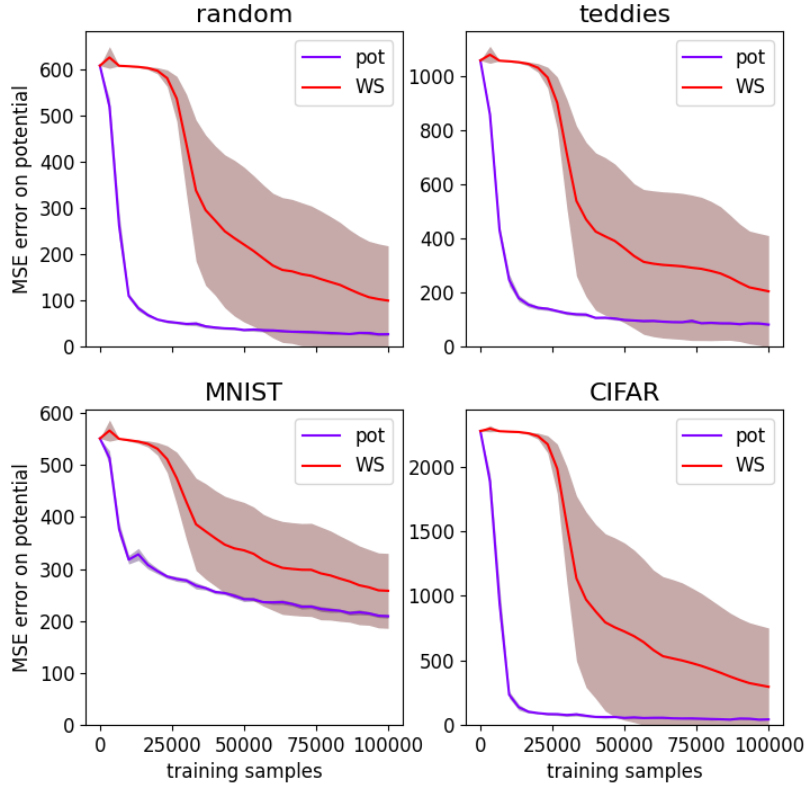


Figure 9: Training with an MSE loss on the potential ('pot') vs. on the negative Wasserstein distance approximation ('WS'). Average over 10 network instances alongside 95% confidence intervals of the mean.³⁰

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, 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.

Why not train on MNIST? The network in [2] was trained on MNIST data. So why don't we? By now, the answer to this question should be clear. Such a network just won't generalize to other datasets, as can be seen in figure 10.

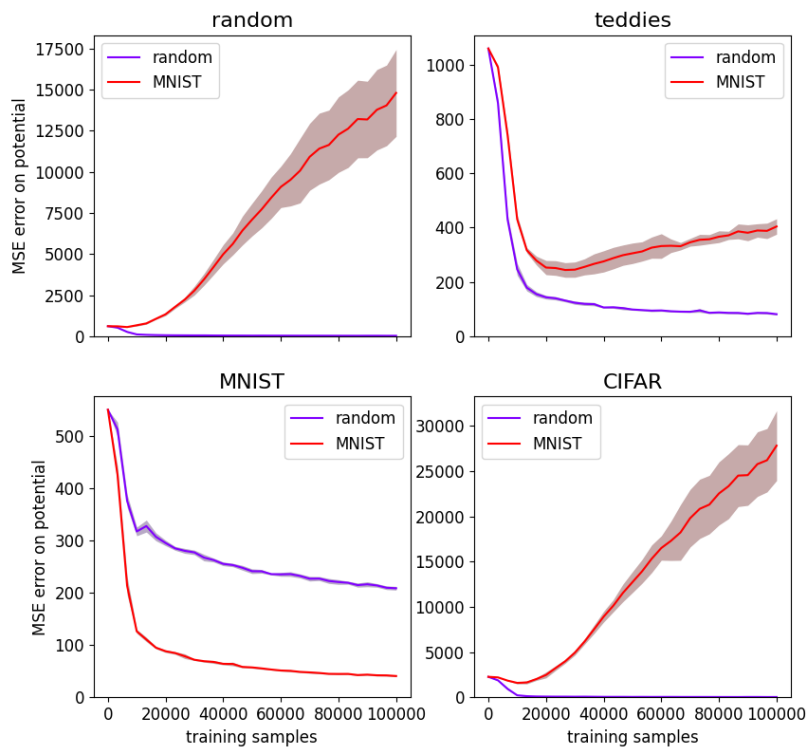


Figure 10: Training on our training data ('random') vs. training on MNIST training data ('MNIST'). Average over 10 samples alongside 95% confidence intervals of the mean.³¹

As expected, the MNIST trained network performs better on the MNIST test dataset, but worse on all other datasets. Interestingly, for the other three datasets, one can notice however that the error for the MNIST trained network became smaller at first – in the 'teddies' case significantly so – before exploding. This shows that the MNIST dataset does contain some useful information for generalization (although limited), but the network quickly starts to overfit on this specific dataset.

³⁰The confidence intervals are shown as shaded areas around the mean. For the 'pot' plots they are very narrow, that is why they are almost not visible. As can be seen from the plots, the confidence intervals for the loss as in equation (22) on the other hand are very wide. This can be explained by a huge discrepancy in performance between the 10 instances, some learning almost as well as with the 'pot' loss, others learning almost nothing at all. Apparently, using this loss sometimes gets the network stuck in very bad local minima.

³¹The MNIST training data has been generated with a constant $k_2 = 0.001$ as in algorithm 3. For every sample, two distributions from the 60.000 available training instances have been chosen at random.

5.6 Training

As can be seen from the confidence intervals in figure 9, different network instances learn at the almost exact same rate when using an error on the dual potential. Hence, for our experiments, we will only consider a single network (as opposed to computing the mean values over multiple networks). Training saturates at around one million training samples, as can be seen in figure 11, with average MSEs on the four test datasets being 11.8, 53.3, 160.7 and 13.4 (for 'random', 'teddies', 'MNIST' and 'CIFAR' respectively), whereas the reported errors without training were 608, 1059, 551 and 2275. In figure 11, the training progress over three million training samples can be seen. The learning rate used decayed from 0.005 to 0.0002.

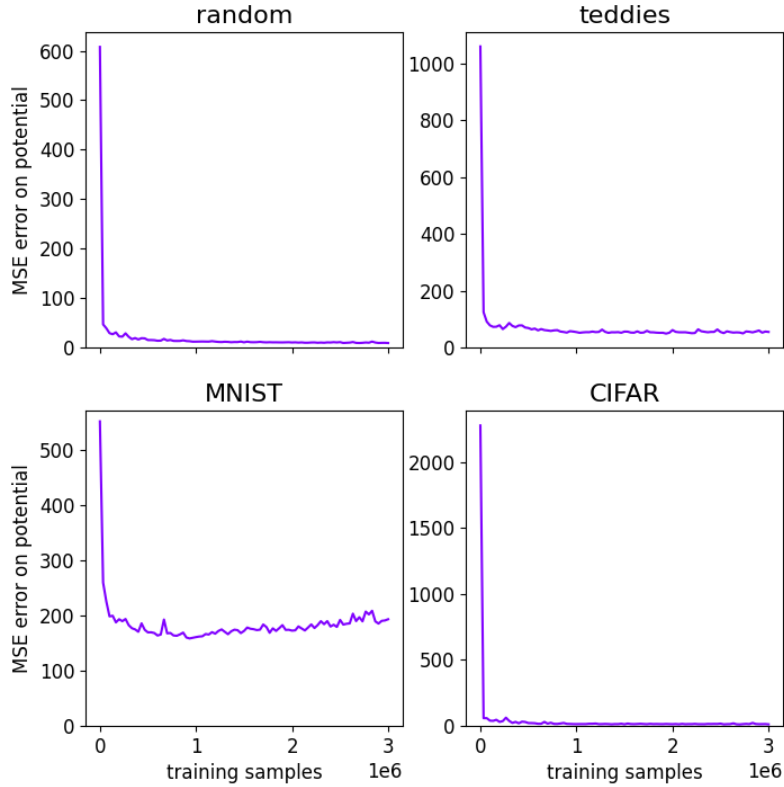


Figure 11: Training a network with $lr = 0.005$ to $lr = 0.0002$ on three million samples.

As can be seen from the plot, most of the learning happens during the first 100,000 samples. In our experiments, we will be using a network trained on one million unique samples, i.e. we only learned for one epoch on the data, as more epochs would not increase performance as can be seen from figure 11. We use a learning rate decaying from 0.005 to 0.0005 and a batch size of 100. Training the network on one million samples takes approximately 100 minutes on a 4 core, 1.6 GHz CPU.

6 Results

In this section, we will see how our Sinkhorn-NN hybrid algorithm performs compared to the Sinkhorn algorithm with its default initialization (i.e. $1_{28} \in \mathbb{R}^{28}$). We used a network trained as outlined in section 5.6. The constants b_1 and b_2 from algorithm 2 were set to $1e-35$ and $1e35$ respectively; they can be chosen to be extremely small resp. extremely large, hence they do not significantly alter the initialization vector. The regularizer used in all experiments was $\varepsilon = 0.2$.³²

6.1 Error w.r.t. Iterations

When it comes to measuring the accuracy of the Sinkhorn algorithm for a specific initialization, there are two errors we will consider: The first is the L^1 error on the Wasserstein distance,³³ i.e. if $(T_i^l)_i \in \mathbb{R}^N$ are the transport costs computed from the Sinkhorn algorithm after l iterations for some N -dimensional batch of test data and $(T_i^{\text{true}})_i \in \mathbb{R}^N$ are the ground truth costs, we compute $L^1(T^l, T^{\text{true}})$. Now when using the Sinkhorn algorithm in practice, the ground truth costs are not available, and a common practice is to consider the marginal constraint violations instead, which are a measure for how close the solution from the algorithm is to an actual transport plan. This is the second error we will be considering. There are different possibilities to measure the marginal constraint violation; in figure 5, for example, we considered $\log(\|1_n^\top \gamma^{(l)} - \nu^\top\|_1)$ with $\gamma^{(l)}$ being the transport plan after l Sinkhorn iterations.

In what follows, we will instead consider

$$\frac{(\|1_m^\top \gamma^{(l)} - \nu^\top\|_1) + (\|\gamma^{(l)} 1_n - \mu\|_1)}{2},$$

which also accounts for the marginal constraint violation on μ . In figure 12, we can see the average L^1 error on the Wasserstein distance for 400 to 2400 Sinkhorn iterations. Figure 13 shows the marginal constraint violations for 400 to 2400 Sinkhorn iterations. In table 2, we report the values for 1, 200, 1000 and 2400 iterations. In each of them, the average over 20 test sets with 50 samples each (i.e. the solution for 50 samples was computed in a parallelized fashion as outlined in section 4.2), alongside the 95% confidence interval of the mean, are reported.

As one can see from the table, the Sinkhorn-NN algorithm reduces the error on the Wasserstein distance and the marginal constraint violation for a given number of iterations by up to 50%, depending on

³² $\varepsilon = 0.2$ is empirically a good choice. In some instances, for smaller regularizing constants the errors even became larger. Also, the smaller ε gets, the larger the number of test samples returning NaN due to numerical inaccuracies in the Sinkhorn algorithm. In the experiments in this section, for the CIFAR dataset up to 24% of samples in the test sets resulted in NaN, in which case the average was computed over the remaining $\geq 76\%$ of samples. For all other test sets, no samples resulted in NaN.

³³We will refer to the transport distances computed as Wasserstein distances, although, to be precise, they are the *squared* Wasserstein-2 distances as discussed in section 5.2.

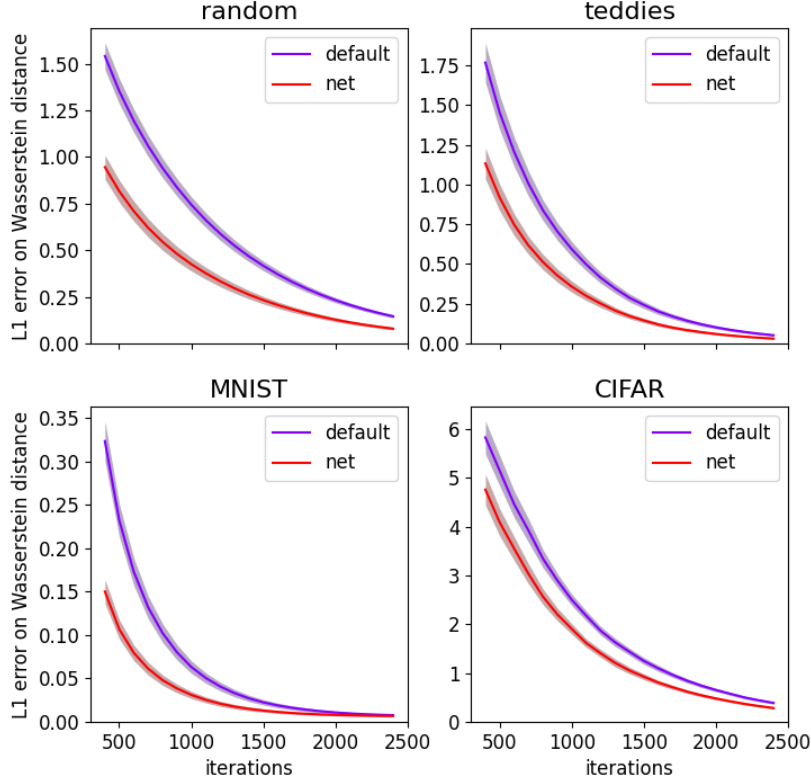


Figure 12: L^1 error on the Wasserstein distance w.r.t. number of Sinkhorn iterations.

the dataset.

In practice, one often uses a threshold on the marginal constraint violation as a stopping criterion. In table 3, we report the average number of iterations needed to achieve a marginal constraint violation of $1e-3$, averaged over 100 individual samples, alongside a 95% confidence interval. The number of iterations needed was measured with an accuracy of 20 iterations (hence, the true values might be up to 20 iterations lower). Table 4 shows the the average number of iterations needed for a marginal constraint violation of $1e-2$.

As one can see from the tables, for various marginal constraint violation thresholds the Sinkhorn-NN hybrid algorithm is consistently outperforming the regular Sinkhorn algorithm. However, the difference becomes more significant for larger thresholds, with an average 33.6% less iterations needed for a $1e-2$ threshold, compared to an average 17.5% less iterations needed for a $1e-3$ threshold.

6.2 Speed

One important aspect is how our algorithm compares against the default Sinkhorn algorithm in terms of time needed for computations. Ultimately, the goal is to compute approximations as quickly as possible. Hence, more important than comparing the errors w.r.t. the number of iterations is com-

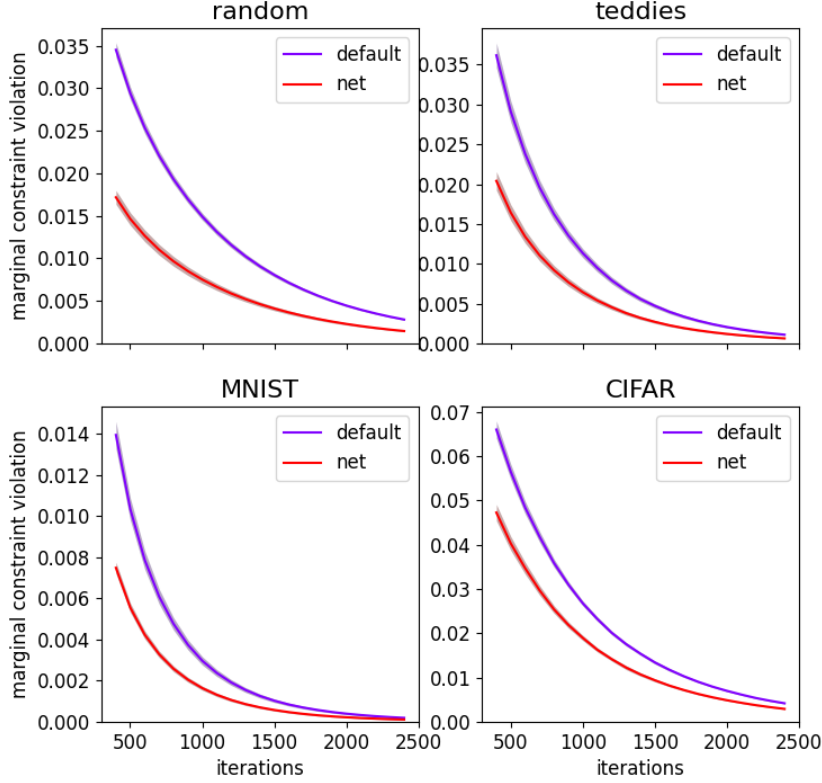


Figure 13: Marginal constraint violation w.r.t. number of Sinkhorn iterations.

paring it with respect to time. Figure 14 shows the time in seconds it took for each of the test subsets (consisting of 50 samples each) from section 6.1 to be computed. The values are averaged over all test subsets and over all four test datasets and presented alongside the 95% confidence intervals of the average.

As one can see from the plot, the Sinkhorn-NN algorithm takes more time for the same number of iterations, as passing the distributions through the network takes additional time, but the difference is negligible. Hence, comparing the two errors – L^1 on the Wasserstein distance and marginal constraint violation – against the time the algorithm took for computations looks very similar to the plots w.r.t. the number of iterations from section 6.1, as can be seen in figures 15 and 16.

Both of them show the respective errors, as usual over 20 test sets with 50 samples each, and the 95% confidence intervals w.r.t. the mean. The time values used were the averages over the 20 test sets. As one can see, the plots are a little more wiggly than the ones w.r.t. the number of iterations, which can probably be explained by time not being as accurately measurable in an objective fashion, as other computer processes in the background might interfere. However, one can still see that the Sinkhorn-NN algorithm achieves errors up to twice as small in the same time as the Sinkhorn algorithm with its default initialization.

	Iterations	random		teddies	
		default	net	default	net
WS	1	5.322 ± 0.083	3.654 ± 0.087	9.856 ± 0.369	7.177 ± 0.383
	200	2.098 ± 0.079	1.338 ± 0.075	2.756 ± 0.163	1.848 ± 0.136
	1000	0.744 ± 0.043	0.425 ± 0.036	0.593 ± 0.047	0.357 ± 0.035
	2400	0.145 ± 0.008	0.079 ± 0.007	0.052 ± 0.006	0.031 ± 0.004
marg	1	$.5869 \pm .0014$	$.5499 \pm .0014$	$.5548 \pm .0047$	$.4790 \pm .0036$
	200	$.0528 \pm .0009$	$.0271 \pm .0008$	$.0611 \pm .0020$	$.0350 \pm .0015$
	1000	$.0149 \pm .0004$	$.0075 \pm .0005$	$.0113 \pm .0006$	$.0064 \pm .0004$
	2400	$.0028 \pm .0001$	$.0015 \pm .0001$	$.0011 \pm .0001$	$.0006 \pm .0001$
		MNIST		CIFAR	
		default	net	default	net
WS	1	10.915 ± 0.397	6.310 ± 0.372	10.011 ± 0.444	8.585 ± 0.443
	200	0.836 ± 0.039	0.387 ± 0.027	7.510 ± 0.397	6.377 ± 0.375
	1000	0.063 ± 0.007	0.031 ± 0.004	2.450 ± 0.118	1.903 ± 0.108
	2400	0.008 ± 0.001	0.007 ± 0.001	0.388 ± 0.024	0.284 ± 0.019
marg	1	$.6242 \pm .0078$	$.4888 \pm .0057$	$.2427 \pm .0043$	$.2412 \pm .0029$
	200	$.0323 \pm .0009$	$.0171 \pm .0005$	$.0957 \pm .0026$	$.0673 \pm .0025$
	1000	$.0030 \pm .0002$	$.0016 \pm .0001$	$.0268 \pm .0006$	$.0190 \pm .0006$
	2400	$.0002 \pm .0000$	$.0001 \pm .0000$	$.0042 \pm .0001$	$.0029 \pm .0001$

Table 2: L^1 error on Wasserstein distance ('WS') and marginal constraint violation ('marg') for 1, 200, 1000 and 2500 Sinkhorn iterations; Sinkhorn-NN hybrid ('net') vs. default initialization ('default').

	random	teddies	MNIST	CIFAR
default	3230 ± 133	2284 ± 115	1310 ± 82	3244 ± 194
net	2520 ± 159	1849 ± 127	1093 ± 73	2842 ± 199

Table 3: Average number of iterations needed to achieve a $1e-3$ marginal constraint violation.

	random	teddies	MNIST	CIFAR
default	1291 ± 84	1025 ± 70	490 ± 37	1569 ± 123
net	721 ± 96	686 ± 71	318 ± 31	1224 ± 133

Table 4: Average number of iterations needed to achieve a $1e-2$ marginal constraint violation.

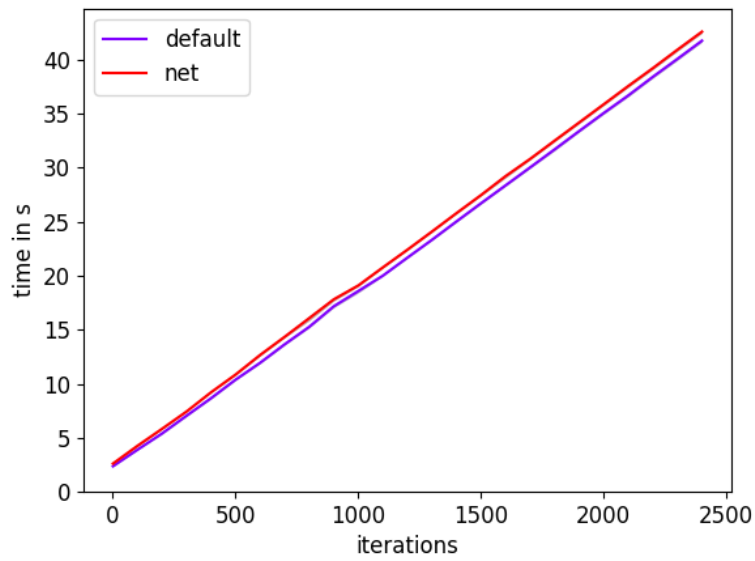


Figure 14: Time in seconds w.r.t. number of Sinkhorn iterations.

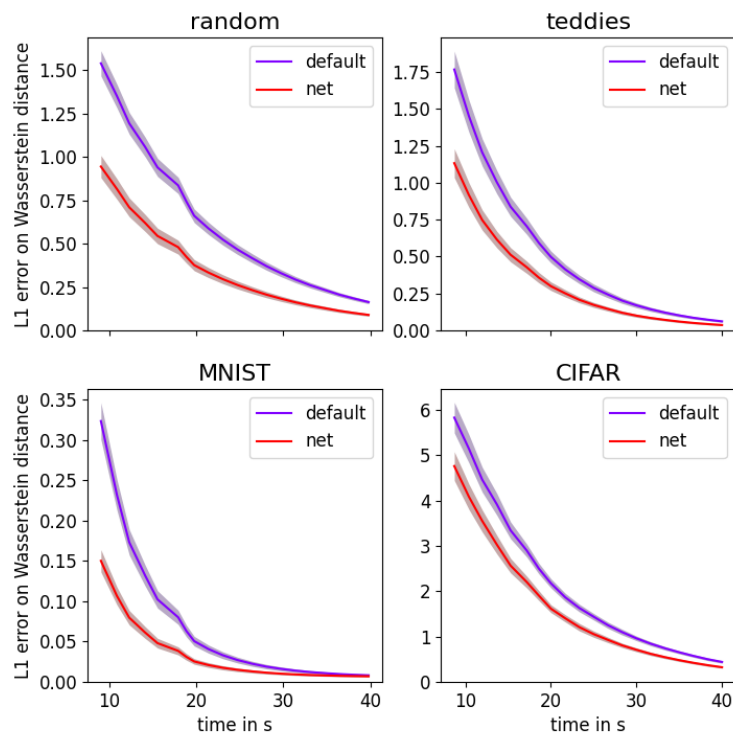


Figure 15: L^1 error on the Wasserstein distance w.r.t. time.

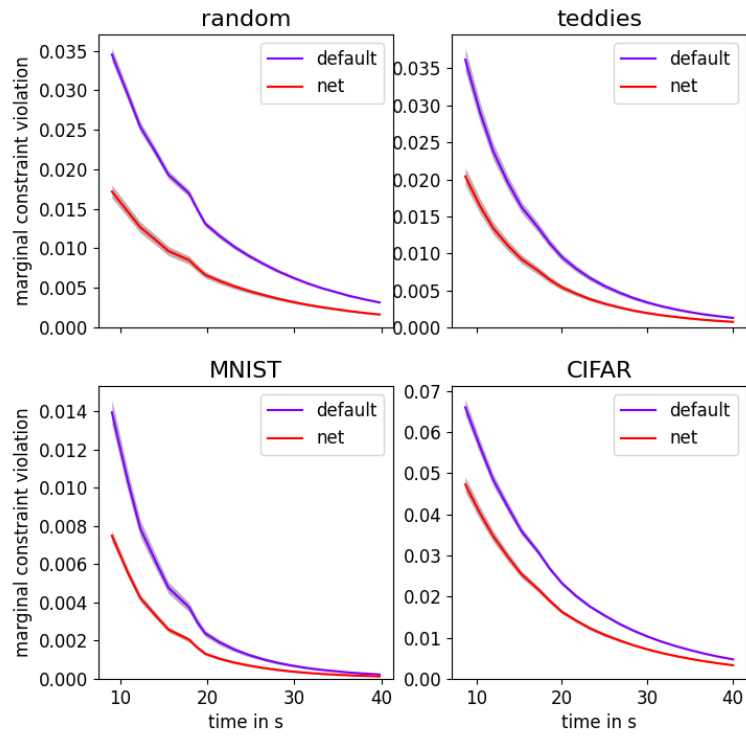


Figure 16: Marginal constraint violation w.r.t. time.

7 Discussion

In this thesis, we explained what the Sinkhorn algorithm is and how and why it works, thoroughly going through the mathematical foundations in chapters 3 and 4, starting with the very basics of optimal transport.

We had a look at how the algorithm is usually initialized, and argued why this initialization might not be advantageous. We then proposed our *Sinkhorn-NN hybrid algorithm*, which features a neural network that learns to predict a potential from the optimal transport dual problem given two distributions, and showed how this network can be used to compute initializations for the Sinkhorn algorithm in chapter 5. We explained how we produce our training data in such a way that it allows for faster learning while still being able to generalize to any data. As can be seen in section 5.5, our loss leads to significantly better approximations of the dual potential than a loss on the Wasserstein distance. Chapter 6 features results from various experiments with the Sinkhorn-NN hybrid algorithm. In section 6.1, we can see that, given a fixed number of iterations for the Sinkhorn algorithm, our hybrid algorithm reduces the error – measured both in terms of the L^1 error on the Wasserstein distance as well as in terms of the marginal constraint violation – by roughly 20% on the CIFAR test dataset and by roughly 40 – 50% on the other test datasets. For achieving a specific threshold on the marginal constraint violation, Sinkhorn-NN consistently outperforms the regular Sinkhorn algorithm; however, the difference grows as the threshold becomes larger. For a $1e-2$ threshold, Sinkhorn-NN needs an average 33.6% less iterations, averaged over all four test datasets, whereas for a $1e-3$ threshold, this difference shrinks to 17.5%. This can be explained by that fact that a good initialization is particularly useful for smaller numbers of iterations, as in this scenario, the default initialization might still be completely off the actual solution. The network also needs time to compute the initialization for the Sinkhorn algorithm, a factor that needs to be taken into account when measuring performance. Hence, in section 6.2, we considered the errors with respect to time needed for computations. The relationship is very similar to the errors with respect to the number of iterations, and we can see that our algorithm achieves errors of up to 50% less, depending on the test dataset and the computation time. This can be explained by our algorithm needing insignificantly more time than the regular Sinkhorn algorithm for the same number of iterations, which is also shown in this section.

These results show that initializing the Sinkhorn algorithm with a vector learned by a neural network can significantly improve its convergence and accuracy. Once such a network is learned, it can be used for all calls of the Sinkhorn algorithm for that particular dimension and cost function.

Further research could include using this approach in much higher dimensions, combining the network with other algorithms such as the Sinkhorn algorithm in the log domain (see [33], chapter 4.4) which also iteratively approximate the dual potentials and need to be initialized with a starting vector, or even using it in the continuous optimal transport scenario, for example with Wasserstein GANs (see, e.g., [3] and [2]). We also noticed that the way in which training data is generated has a huge impact on the performance; e.g. simply altering the constant k_1 in algorithm 3 made a big difference. However, entirely different ways of generating training data are conceivable, and future research could include coming up with a more sophisticated approach.

A Appendix

In this section we recall some definitions and properties that are used throughout the thesis, in particular in chapter 3 on optimal transport. For measure theoretic definitions not provided here, one might resort to introductions to measure theory such as the one in [17]. A great overview of optimal transport in the general setting can be found in [41], and an introduction to discrete optimal transport in [33].

Definition A.1 (σ -Algebra). *Let Ω be a non-empty set. A subset $\mathcal{A} \subset 2^\Omega$ of the power set of Ω is called a σ -algebra if it satisfies the following properties:*

1. $\Omega \in \mathcal{A}$
2. If $A \in \mathcal{A}$, then also $\Omega \setminus A \in \mathcal{A}$
3. If $(A_i)_{i \in \mathbb{N}} \subset \mathcal{A}$, then also

$$\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}$$

Definition A.2 (Borel σ -Algebra). *Let $(\mathcal{T}, \mathcal{O})$ be a topological space. Then the σ -algebra generated by \mathcal{O} is called the Borel σ -algebra on \mathcal{T} .³⁴*

Definition A.3 (Measurable Space, Measure, Finite Measure, Discrete Measure, Signed Measure). *A measurable space is a pair $(\mathcal{X}, \mathcal{A})$ of a non-empty set \mathcal{X} and a σ -algebra \mathcal{A} on \mathcal{X} . Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. A measure on $(\mathcal{X}, \mathcal{A})$ is a map $\mu : \mathcal{A} \rightarrow [0, \infty]$ 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 $\nu : \mathcal{A} \rightarrow [-\infty, \infty]$ 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.4 (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*

³⁴This means that it is defined as the intersection over all σ -algebras containing all sets $O \in \mathcal{O}$, which can be shown to again be a σ -algebra.

sets \mathcal{X}_μ and \mathcal{X}_ν 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.5 (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 [16], theorem 2. □

Definition A.6 (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.7 (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 $\mathcal{A} \otimes \mathcal{B}$.*

Definition A.8 (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.9 (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 [28], chapter 8.2, theorem A. □

Definition A.10 (Product Topology). *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 $\mathcal{T}_{\mathcal{X} \times \mathcal{Y}}$.*

Remark A.11 (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: 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 $(\mathcal{X} \times \mathcal{Y}, \mathcal{T}_{\mathcal{X} \times \mathcal{Y}})$. 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.³⁵ 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.³⁶

³⁵ A proof of this result can e.g. be found in [23], lemma 1.2. Note that completeness of the spaces is not even needed.

³⁶ Here we make a slight abuse of notation, as our definition of a Polish space 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.

Lemma A.12. *Let (\mathcal{X}, d) be a metric space and $x, y \in \mathcal{X}$. Then for any $z \in \mathcal{X}$ there holds*

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

Proof. We have

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

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

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

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. □

Lemma A.13. *Let (\mathcal{X}, d) be a metric space. Then the following are equivalent:*

- (i) \mathcal{X} is compact
- (ii) \mathcal{X} is sequentially compact, i.e. any sequence in \mathcal{X} contains a convergent subsequence
- (iii) \mathcal{X} is complete and totally bounded

Proof. (i) \Rightarrow (ii) :

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.

(ii) \Rightarrow (iii) :

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.

(iii) \Rightarrow (i) :

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 $C_1^1, \dots, C_{p_1}^1$ 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 $C_1^2, \dots, C_{p_2}^2$ of radius less than $\frac{1}{2}$ (without loss of generality let $C_i^2 \subset C^1$ for all $i \in \llbracket p_2 \rrbracket$), and again one of them, C^2 , cannot be covered by finitely many U_i . Proceeding like this, we find a sequence $C^1 \supset C^2 \supset C^3 \supset \dots$ 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 . \square

Lemma A.14. *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. As K is closed, the restriction of (\mathcal{X}, d) to K is still a complete, metric space. Now the statement follows immediately from lemma A.13. \square

Theorem A.15 (Weierstraß). *Let (\mathcal{X}, d) be a metric space, $K \subset \mathcal{X}$ compact, and $I : \mathcal{X} \rightarrow [-\infty, \infty]$ 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 [17], as the statement there is a generalization of our proposition to topological spaces. \square

Theorem A.16 (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 [9]. \square

Theorem A.17 (Disintegration). *Let \mathcal{X} and \mathcal{Y} be two Polish spaces and $\mu \in P(\mathcal{X})$. Let $\pi : \mathcal{X} \rightarrow \mathcal{Y}$ 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*

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

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

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

Proof. A proof to this theorem can be found in [13], Chapter III, theorem 70 on page 78. It holds true in the even more general setting of Radon spaces. \square

Theorem A.18 (Monotone Convergence). *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and $(f_n)_{n \in \mathbb{N}} : \mathcal{X} \rightarrow [c, \infty]$ 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

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

Proof. A proof can e.g. be found in [8], Theorem 2.8.2. \square

Lemma A.19 (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 $\int_{\mathcal{X}} \varphi \, d\mu = \int_{\mathcal{X}} \varphi \, d\nu$ for all $\varphi \in C_b(\mathcal{X})$.*

Proof. If $\mu = \nu$, then $\int_{\mathcal{X}} \varphi \, d\mu = \int_{\mathcal{X}} \varphi \, d\nu$ for all $\varphi \in C_b(\mathcal{X})$ is clear. Now let $\int_{\mathcal{X}} \varphi \, d\mu = \int_{\mathcal{X}} \varphi \, d\nu$ hold for all $\varphi \in C_b(\mathcal{X})$. Let $A \in \mathcal{B}(\mathcal{X})$ be closed and for $\varepsilon > 0$ define $f_\varepsilon : \mathcal{X} \rightarrow [0, 1]$, $f_\varepsilon(x) = \max\{1 - \frac{1}{\varepsilon}d(x, A), 0\}$ (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 as $\varepsilon \rightarrow 0$. Hence, by proposition A.18, $\mu(A) = \nu(A)$. \square

Proposition A.20 (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*

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

Proof. A proof of this famous inequality can e.g. be found in [8], theorem 2.11.9. \square

Proposition A.21 (Lagrange Multipliers). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i \in \llbracket m \rrbracket$ be continuously differentiable. Consider the minimization problem*

$$\begin{aligned} & \min_x f(x) \\ & \text{s.t. } h_i(x) = 0 \text{ for all } i \in \llbracket m \rrbracket. \end{aligned}$$

Let x^ be a local minimum of this problem, such that all $\nabla h_i(x^*)$ for $i \in S \subset \llbracket m \rrbracket$ are linearly independent. Then there exists a unique vector $\lambda \in \mathbb{R}^{|S|}$, where $|S|$ denotes the number of elements in S , called the Lagrange multiplier, such that*

$$\nabla f(x^*) + \sum_{i \in S} \lambda_i \nabla h_i(x^*) = 0.$$

Proof. This very well-known statement from calculus is e.g. proven in [19], theorem 1.13. \square

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] M. Bacharach. *Estimating nonnegative matrices from marginal data*. International Economic Review, 6(3):294–310, 1965.
- [5] D. P. Bertsekas. *Nonlinear Programming*. 2nd edition, Athena Scientific, Belmont, Massachusetts, 1999.
- [6] D. P. Bertsekas. *A new algorithm for the assignment problem*. Mathematical Programming, 21(1):152–171, 1981.
- [7] D. Bertsimas, J. Tsitsiklis. *Introduction to Linear Programming*. Athena Scientific and Dynamic Ideas, Belmont, Massachusetts, 1997.
- [8] V. Bogachev. *Measure Theory Volume I*. Springer, Berlin, 2007.
- [9] V. Bogachev. *Measure Theory Volume II*. Springer, Berlin, 2007.
- [10] N. Courty, R. Flamary, A. Habrard, A. Rakotomamonjy. *Joint distribution optimal transportation for domain adaptation*. Advances in Neural Information Processing Systems, 30, 2017.
- [11] M. Cuturi. *Sinkhorn distances: lightspeed computation of optimal transport*. Advances in Neural Information Processing Systems 26, pages 2292–2300, 2013.
- [12] R. Dadashi, L. Hussenot, M. Geist, O. Pietquin. *Primal Wasserstein Imitation Learning*. arXiv:2006.04678 [cs.LG], 2020.
- [13] C. Dellacherie, P.-A. Meyer. *Probabilities and Potential*. Hermann, Paris, 1978.
- [14] E. Deming, F. F. Stephan. *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*. Annals of Mathematical Statistics, 11(4):427–444, 1940.
- [15] S. Erlander. *Optimal Spatial Interaction and the Gravity Model*, volume 173. Springer-Verlag, 1980.
- [16] T. Fischer. *Existence, uniqueness, and minimality of the Jordan measure decomposition*. arXiv:1206.5449 [math.ST], 2012.
- [17] I. Fonseca, G. Leoni. *Modern Methods in the Calculus of Variations: L^p Spaces*. Springer, New York, 2007.

- [18] J. Franklin, J. Lorenz. *On the scaling of multidimensional matrices*. Linear Algebra and its Applications, 114:717–735, 1989.
- [19] A. Fuente. *Mathematical Methods and Models for Economists*. Cambridge University Press, Cambridge, 2000.
- [20] A. Galichon, B. Salanié. *Matching with trade-offs: revealed preferences over competing characteristics*. Technical report, Preprint SSRN-1487307, 2009.
- [21] A. Galichon. *Optimal transport methods in economics*. Optimal Transport Methods in Economics, Princeton University Press, 2016.
- [22] S. Ioffe, C. Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. arXiv:1502.03167 [cs.LG], 2015.
- [23] O. Kallenberg. *Foundations of Modern Probability*. 2nd edition, Springer, New York, 2002.
- [24] L. V. Kantorovich. *On the Translocation of Masses*. Dokl. Akad. Nauk SSSR, 37, No. 7–8, 227–229, 1942, available here in an English translation (last visited 10/12/2022).
- [25] D. P. Kingma, J. Ba. *Adam: A Method for Stochastic Optimization*. arXiv:1412.6980 [cs.LG], 2017.
- [26] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, G. K. Rohde. *Optimal mass transport: Signal processing and machine-learning applications*. IEEE signal processing magazine, 34(4):43–59, 2017.
- [27] H. W. Kuhn. *The hungarian method for the assignment problem*. Naval Research Logistics Quarterly, 2:83–97, 1955.
- [28] M. Loève. *Probability Theory I*. 4th edition, Springer, 1977.
- [29] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Royale Sci. Paris, 3, 1781.
- [30] L. Nenna. *Lecture 2: Entropic Optimal Transport*. Available online: <https://lucanenna.github.io/teaching/optimaltransport/lecture2.pdf>, 2022. Last visited 10/11/22.
- [31] J. B. Orlin. *A polynomial time primal network simplex algorithm for minimum cost flows*. Mathematical Programming, 78(2):109–129, 1997.
- [32] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press Inc., 1967.
- [33] G. Peyré, M. Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, vol. 11, number 5-6, 355–607, 2019.
- [34] A. Pratelli. *On the equality between Monge’s infimum and Kantorovich’s minimum in optimal mass transportation*. Elsevier Masson SAS, 2006.
- [35] F. Santambrogio. *Optimal transport for applied mathematicians*. Birkhäuser, 2015.
- [36] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. *Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming*. Cell, 176(4), 2019.

-
- [37] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, J.-L. Starck. *Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning*. SIAM Journal on Imaging Sciences, 11(1):643–678, 2018.
 - [38] R. Sinkhorn, P. Knopp. *Concerning nonnegative Matrices and doubly stochastic Matrices*. Pacific Journal of Mathematics, Vol. 21, No. 2, 1967.
 - [39] R. E. Tarjan. *Dynamic trees as search trees via euler tours, applied to the network simplex algorithm*. Mathematical Programming, 78(2):169–177, 1997.
 - [40] J. Thornton, M. Cuturi. *Rethinking Initialization of the Sinkhorn Algorithm*. arXiv:2206.07630v1 [stat.ML], 2022.
 - [41] C. Villani. *Optimal Transport Old and New*. Springer, Berlin Heidelberg, 2009.
 - [42] A. G. Wilson. *The use of entropy maximizing models, in the theory of trip distribution, mode split and route split*. Journal of Transport Economics and Policy, pages 108–126, 1969.
 - [43] G. U. Yule. *On the methods of measuring association between two attributes*. Journal of the Royal Statistical Society, 75(6):579–652, 1912.

Index

- $(\mathcal{P}_p(\mathcal{X}), W_p)$, 30
- 1-Lipschitz, 19
- $B_r(x)$, 9
- $C(\mathcal{X})$, 9
- $C_b(\mathcal{X})$, 9, 66
- H , 35
- L^1 , 10
- $P(\mathcal{X})$, 9
- S_n , 8
- W_1 , 29
- W_p , 28
- $\Pi(\mu, \nu)$, 12
- δ_x , 9
- $\langle \cdot, \cdot \rangle$, 8
- $\llbracket m, n \rrbracket$, 8
- $\llbracket n \rrbracket$, 8
- $\mathcal{A} \otimes \mathcal{B}$, 63
- $\mathcal{L}(X)$, 10
- $\mathcal{L}(X, Y)$, 10
- $\mathcal{P}_p(\mathcal{X})$, 28
- $\mathcal{T}_{\mathcal{X} \times \mathcal{Y}}$, 63
- $\mu \otimes \nu$, 63
- μ^+ , 63
- μ^- , 63
- $\partial^c \psi$, 20
- $\partial^c \psi(x)$, 20
- $\pi_{\mathcal{X}}$, 9
- ψ^c , 18
- σ -algebra, 62
- σ -finite, 63
- Id , 9
- $\text{vec}(A)$, 8
- φ^c , 18
- c -concavity, 19, 45, 50
- c -convexity, 20
- c -cyclical monotonicity, 22
- c -subdifferential, 20
- c -superdifferential, 20
- c -transform, 18, 45, 50, 52
- Adam optimizer, 49
- algorithm
 - auction, 34
 - Hungarian, 34
 - network simplex, 34
 - simplex, 34
 - Sinkhorn, 34, 41, 45, 55, 61
 - convergence, 42
 - initialization, 43, 45
 - parallelization, 43
 - Sinkhorn-NN hybrid, 45, 55, 61
- batch normalization, 49
- batch size, 54
- Borel σ -algebra, 62
- Borel σ -algebra on product space, 63
- CIFAR10, 48, 49, 52–58, 60, 61
- competitiveness, 18
- cost matrix, 48
- coupling, 11
 - deterministic, 12
 - trivial, 12, 35
- domain adaptation, 6
- duality
 - entropic optimal transport, 39
 - optimal transport, 27
- Earth Mover’s distance, 28
- entropy, 35
- epoch, 54
- Frobenius dot-product, 8
- function
 - strongly convex, 36
- Gibbs kernel, 37
- gluing lemma, 29

- imaging, 6
- imitation learning, 6
- Jordan decomposition, 63
- Kantorovich, 6
 - dual problem, 17
 - problem, 13, 32
- Kantorovich-Rubinstein distance, 28
- Kronecker product, 8
- Lagrange multiplier, 37, 38, 66
- Lagrangian, 38
- law, 10
- learning rate, 10, 49, 54
- linear program
 - dual, 33
 - optimal solution, 33
 - primal, 33
- lower semicontinuity, 9
 - of the cost functional, 15
- lr, 10, 54
- marginal, 11
 - constraint violation, 41, 55, 58, 60, 61
- measurable space, 62
- measure, 62
 - Borel, 9
 - concentrated, 9
 - Dirac, 9
 - discrete, 62
 - finite, 9, 62
 - mutually singular, 62
 - product, 63
 - pushforward, 10
 - signed, 23, 62
 - support, 9
- measure space, 62
- Minkowski inequality, 32, 66
- MNIST, 47–49, 52–58, 60
- Monge, 6
 - problem, 12
- MSE, 10, 49
- negligible, 9
- neighbourhood, 9
- network architecture, 49
- optimal transport, 11, 61
 - discrete, 32
 - dual, 33, 45
 - primal, 32, 45
 - dual problem, 17, 51
 - entropic, 35
 - dual, 39, 45, 51
 - primal, 36, 45
 - solution, 37
 - primal problem, 13
- Polish space, 9, 63
- potential, 18
- precompact, 65
- product σ -algebra, 63
- product topology, 63
- Quick, Draw!, 48, 49, 52–54, 56–58, 60
- regularizer, 36, 51
- ReLU, 49
- sequentially compact, 64
- signal processing, 6
- Sinkhorn-Knopp fixpoint iteration, 41
- support
 - of a function, 9
 - of a measure, 9
- test data, 48
- theorem
 - disintegration, 30, 65
 - duality, 27, 45, 50–52
 - fundamental theorem of optimal transport, 23
 - monotone convergence, 16, 65
 - Prokhorov, 14, 65
 - Weierstraß, 15, 65
- tightness
 - of a measure, 14
 - of a set, 10
 - of prices, 18
 - of transport plans, 14
- totally bounded, 9

training data, 46, 61

transport map, 12

transport plan, 11, 37, 39, 51, 55

 optimal, 13

 existence, 16

variation

 lower, 23, 63

 total, 63

 upper, 63

Wasserstein, 28

 barycenter, 28

 distance, 28, 37, 48, 50, 61

 is a metric, 30, 37, 50

 space, 28

 is Polish, 32

weak convergence, 10

weak topology, 10