The Optimal Transport Problem

Master Thesis

Oscar Ramirez

The Optimal Transport Problem

Master Thesis

by

Oscar Ramirez

to obtain the degree of Master of Science
in Mathematical Modelling and Engineering,
to be defended publicly on September, 2018.

Project duration: September, 2016 – September, 2018
Thesis committee: Prof. Juan Enrique Martinez Legaz, UAB, supervisor

Preface

Preface…

Oscar Ramirez
Barcelona, September 2018

Contents

Notation Table.

$\emptyset$	Empty set
$\mathbb{R}$	Real numbers field.
$\overline{\mathbb{R}}$	$\mathbb{R} \cup \{+\infty\}$. That is $[-\infty, \infty]$
$\mathbb{R}_+$	The set of nonnegative real numbers, that is the interval $[0, \infty)$.
$\overline{\mathbb{R}}_+$	The set of nonnegative extended real numbers, that is the interval $[0, \infty]$
δ_x	The Dirac mass at point x.
$\mathbb{R}^d$	The d-dimensional Euclidean space.
id	Identity map.
$\mathcal{M}(X)$	Space of measures on X.
$\mathcal{M}_+(X)$	Space of positive measures on X.
$\mathcal{P}(X)$	Space of probabilities on X.
$\mu \ll \nu$	The measure is absolutely continuous with respect to ν.
$\mathbb{1}_\Omega$	Indicator function of a set Ω. If $x \in \Omega$ then $\mathbb{1}_\Omega(x) = 1$. If $x \in \Omega^c$, we have $\mathbb{1}_\Omega(x) = 0$.
$\mu \llcorner A$	A measure μ restricted to a set A.
ω_d	The Measure of the unite ball in $\mathbb{R}^d$.
$\wedge$	The min operator, that is $a \wedge b := \min\{a, b\}$.
$\vee$	The max operator, that is $a \vee b := \max\{a, b\}$.
$\mathrm{D}T(x)$	Jacobian matrix of a map $T(x)$.
$T_\# \mu$	The image measure (or pushforward measure) of μ through the map T.
$f_{\mid\Omega}$	The restriction of a function f to a set Ω.
$\Pi(\mu, \nu)$	The set of transport plans from μ to ν.
$\frac{\delta F}{\delta \rho}$	First variation of $F : \mathcal{P}(X) \to \mathbb{R}$, that is $\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\rho + \epsilon\chi)\Big\vert_{\epsilon=0} = \int \frac{\delta F}{\delta \rho} \mathrm{d}\chi$
W_p	Wasserstein distance of order p.
$\mathbb{W}_p$	Wasserstein space of order p.
γ_T	The transport plan in $\Pi(\mu, \nu)$ induced by a map T. That is $\gamma_T = (\mathrm{id}, T)_\# \mu$ and $T_\# \mu = \nu$.
$M(T)$	Monge cost of a map T.
$K(\gamma)$	Kantorovich cost of a plan γ.
$\mu \otimes \nu$	The product measure of μ and ν such that $\mu \otimes \nu(A \times B) = \mu(A)\nu(B)$.
$M^{k \times h}$	The set of real matrices with k rows and h columns.
$M^\top$	Transpose of a matrix M.
i.i.d.	Independent and identical probability distributions.
l.s.c.	Lower semicontinuous.
$\mathscr{L}^p$	Lebesgue measure on $\mathbb{R}^p$
$\mathscr{H} \llcorner A$	Hausdorff measure applied to some set $A \subset \mathbb{R}^d$.
$B(x, \epsilon)$	Open ball with radius ϵ centered at x.
$\overline{B}(x; \epsilon)$	Closed ball with radius ϵ centered at x.

1

Preliminaries.

We start this chapter reminding the basic definitions and theorems in topology and measure theory, since they are needed to have a suitable framework to discuss the optimal transport problem and its applications.

Definitions and important theorems to remember.

Topology.

We start with the definition of topology that is needed to introduce a notion of continuity. We refer to [3] for more details in topology.

Definition 1.1 (Topology)**.** *A topology on a set X is a collection $\mathcal{T}$ of subsets of X having the following properties*

- *The space X itself and $\emptyset$ are in $\mathcal{T}$.*
- *The union of the elements of any sub-collection of $\mathcal{T}$ is in $\mathcal{T}$.*
- *The intersection of the elements of any finite sub-collection of $\mathcal{T}$ is in $\mathcal{T}$.*

A pair $(X, \mathcal{T})$ is called a topological space. The elements of $\mathcal{T}$ are called open sets. The complements of the sets of $\mathcal{T}$ are called closed sets. The interior of a set A, is defined as the biggest open set contained in A. Similarly, the closure of a set A, is defined as the smallest closed set containing A. We use indistinctly the notation $\mathrm{int}(A)$ and A° for the interior of a set A. In the same way, for the closure we use the notation $\mathrm{clo}(A)$ or $\bar{A}$. An equivalent way to define the same ideas is give by the following,

$$\mathrm{int}(A) = \bigcup_{B \text{ is open.}} B$$

$$\mathrm{clo}(A) = \bigcap_{B \text{ is closed.}} B$$

We remark that a set is open if and only if $A = A^{\circ}$, and a set is closed if and only if $A = \bar{A}$. We call a neighborhood of x an element of $\mathcal{T}$ containing x.

Definition 1.2 (Topological Basis.)**.** *Give a set X endowed with a topology $\mathcal{T}$. We call a basis for $\mathcal{T}$ is a collection $\mathcal{B}$ of subsets of X (called basis elements), such that,*

1. *For each $x \in X$, there is at least one basis element B containing x.*
2. *If x belongs to the intersection of two basis elements B_1 and B_2, then there is a basis element B_3 containing x such that $B_3 \in B_1 \cap B_2$.*

Definition 1.3 (Dense set). *A subset D of a topological space X is dense in X if for any point x in X, any neighborhood of x contains at least one point from D.*

Definition 1.4 (Separable space). *A topological space is called separable if it contains a countable, dense subset.*

Topologies in which one element is not a closed set, or in which a sequence can converge to more than one point, are not really interesting for practical problems. If such things are allowed the theorems that one can prove are limited. A mathematician Felix Hausdorff suggested to add the following condition:

Definition 1.5 (Hausdorff space). *A topological space X is called a Hausdorff space if for each pair x_1, x_2 of distinct points of X, there exist neighborhoods U_1, and U_2 of x_1 and x_2, respectively, that are disjoint.*

Definition 1.6 (Distance).

Definition 1.7 (Metric Spaces).

Definition 1.8 (Completeness). *A metric space X is called complete if every Cauchy-Sequence of points in X has a limit that is also in X.*

Definition 1.9 (Completely metrizable space).

There is a subtle difference between complete metric space and completely metrizable space. And the difference lies on the words "*there exists at least a metric...*" in the completely metrizable definition, and "*given a metric*". Complete metrizable is a topological property while completeness is a property of the chosen metric.

Definition 1.10 (Polish space). *We call Polish space to any topological space that is separable and completely metrizable.*

Definition 1.11 (Sequentially compact). *A subset K of a metric space X is said to be compact if from any sequence x_n, we can extract a converging subsequence $x_{n_k} \to x \in K$.*

Definition 1.12 (Compactness). *A subset K of a metric space X is compact if every open cover of K has a finite subcover.*

Theorem 1.1. *A subset of a metric space is compact if and only if it is sequentially compact.*

Definition 1.13 (Liminf and Limsup). *Let X be a Hausdorff space. Let $\mathcal{V}(x_0)$ be a topological basis of X, such that all $V \in \mathcal{V}$ contains x_0. Let $f : X \to \overline{\mathbb{R}}$ a functional valued in $\overline{\mathbb{R}}$. We define,*

$$\liminf_{x \to x_0} f(x) = \sup_{V \in \mathcal{V}(x_0)} \inf_{s \in V} f(s)$$

$$\limsup_{x \to x_0} f(x) = \inf_{V \in \mathcal{V}(x_0)} \sup_{s \in V} f(s)$$

The above definitions can be expressed in terms of sequences of real numbers. Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in X, the above formulation is equivalent to say.

$$\liminf_{n \in \mathbb{N}} x_n := \lim_{n \to \infty} \left(\inf_{m \geq n} x_m \right)$$

Equivalently for $\limsup$,

$$\liminf_{n \in \mathbb{N}} x_n := \lim_{n \to \infty} \left(\sup_{m \geq n} x_m \right)$$

Please note that the convergence to some point x_0, $(x_n)_{n \in \mathbb{N}} \to x_0$ is not required in the last definitions.

Functional Analysis.

Definition 1.14 (Linear Space).

Definition 1.15 (Banach Space).

Definition 1.16 (Inner product).

Definition 1.17 (Hilbert Space).

Definition 1.18 (Continuity).

For a given be a metric space X . We denote the set of continuous, real-valued functions $f : X \to \mathbb{R}$ by $C(X)$.

Theorem 1.2. *Let $K \subset X$ a compact subset of a metric space X. The space $C(K)$ is complete.*

A natural norm on spaces of continuous functions is the uniform norm (also called infinity norm), which is defined by,

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

The norm $||f||_\infty$ is finite if and only if f is bounded. And we use $C_b(X)$ to refer the space of bounded functions on X.

Definition 1.19. *Let f be a real valued function, $f : X \to \mathbb{R}$ on a metric space. The **support of a function**, $\operatorname{supp} f$ is the closure of the set on which f is nonzero.*

$$\operatorname{supp} f = \operatorname{clo}(\{x \in X : \quad f(x) \neq 0\})$$

We say that f has compact support if $\operatorname{supp} f$ is a compact subset of X, and denote the space of continuous functions on X with compact support by $C_c(X)$.

The space $C_c(X)$ is a linear subspace of $C_b(X)$, but it does not need to be closed.

Definition 1.20. *Suppose that X is a separable and locally compact metric space. We say that a real valued function f belongs to $C_0(X)$ if and only if $f \in C(X)$, and for every $\epsilon > 0$, there exists a compact set $K \subset X$ such that $|f| < \epsilon$ on $X \setminus K$.*

Definition 1.21. *Let $\mathcal{F}$ be a family of functions from a metric space (X, d) to a metric space (Y, d). The family $\mathcal{F}$ is equicontinuous if for every $x \in X$ and $\epsilon > 0$ there is $\delta > 0$ such that $d(x, y) < \delta$ implies $d(f(x), f(y)) < \epsilon$ for all $f \in \mathcal{F}$.*

Theorem 1.3. *An equicontinuous family of functions from a compact metric space to a metric space is uniformly equicontinuous.*

Theorem 1.4 (Ascoli-Arzelà). *Let K be a compact metric space. A subset M of the set of continuous functions $M \subset C(K)$ is compact if and only if it is closed, bounded and equicontinuous. That is any sequence $(f_n)_{n \in \mathbb{N}}$ in M admits a subsequence converging to f in M.*

Definition 1.22 (Proper convex function). *Let $f : X \to \overline{\mathbb{R}}$, a function taking values in the extended real number line. We call it proper convex function if $\exists x \in X$ such that $f(x) < \infty$. And $\forall x \in X$, $f(x) > -\infty$.*

Definition 1.23 (Projection of a Cartesian Product.). *Let $\operatorname{proj}_x : X \times Y \to X$ be defined by the equation*

$$\operatorname{proj}_x(x, y) = x;$$

Equivalently, let $\operatorname{proj}_y : X \times Y \to Y$ be defined by,

$$\operatorname{proj}_y(x, y) = y$$

These maps proj_x and proj_y are called the projections of $X \times Y$ onto X and Y respectively.

We can generalize a definition over a general Cartesian product. Given a set X, we define a *J-tuple* of elements of X to be a function $\mathbf{x} : J \to X$. If α is an element of J, we often denote the value of $\mathbf{x}$ at α by x_α rather than $\mathbf{x}(\alpha)$; we call it the α-th coordinate of $\mathbf{x}$. And we often denote the function $\mathbf{x}$ by the symbol.

$$(x_\alpha)_{\alpha \in J}$$

Let $\{A_\alpha\}_{\alpha \in J}$ be a set of indexed family of sets; let $X = \bigcup_{\alpha \in J} A_\alpha$. The *Cartesian product* of this indexed family, denoted by

$$\prod_{\alpha \in J} A_\alpha$$

is defined to be the set of all J-tuples $(x_\alpha)_{\alpha \in J}$ of elements of X such that $x_\alpha \in A_\alpha$ for each $\alpha \in J$. That is, it is the set of all functions,

$$\mathbf{x} : J \to \bigcup_{\alpha \in J} A_\alpha$$

such that $\mathbf{x}(\alpha) \in A_\alpha$ for each $\alpha \in J$.

Definition 1.24 (Lower Semicontinuity)**.** *On a complete metric space X, a function $f : X \to \overline{\mathbb{R}}$ is said to be lower semi-continuous (l.s.c.) if for every sequence $(x_n)_{n\in\mathbb{N}}$ converging to $x \in X$, we have*

$$f(x) \leq \liminf_{n\in\mathbb{N}} f(x_n)$$

We can see from the above definition that any continuous function is lower-semicontinuous. In other words, lower-semicontinuity is a milder requirement than continuity, although it preserves interesting properties that can be exploited in optimization.

Proposition 1.1. *Let $f : X \to \overline{\mathbb{R}}$ be a convex and lower-semicontinuous function. Assume that there exists $x_0 \in X$ such that $f(x_0) = -\infty$. Then f is nowhere finite on X.*

Theorem 1.5. *If f_α is an arbitrary family of lower semi-continuous functions on X, then $f = \sup_\alpha f_\alpha$ is also lower-semicontinuous.*

Definition 1.25 (Lipschitz condition)**.**

Theorem 1.6. *Let $f : X \to \overline{\mathbb{R}} \backslash -\infty$ be a function bounded from below. Then f is l.s.c. if and only if there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of k-Lipschitz functions such that for every $x \in X$, $f_n(x)$ converges increasingly to $f(x)$.*

Definition 1.26. *Let $F : X \to \overline{\mathbb{R}}$ be a given functional bounded from below on a metric space X. Let $\mathcal{G}$ be the set of lower semicontinuous functions $G : X \to \overline{\mathbb{R}}$, such that $G \leq F$. We call a relaxation the supremum of $\mathcal{G}$. This functional does exist since the supremum of an arbitrary family of lower semicontinuous functions is also lower semicontinuous. It is possible to have a representation formula as follows:*

$$\bar{F}(x) = \inf\left\{\liminf_{n\in\mathbb{N}} F(x_n) : \ x_n \to x\right\}. \tag{1.1}$$

As consequence of this definition we see that $F \geq \bar{F}$ implies $\inf F \geq \inf \bar{F}$. Let $l = \inf F$ then $F \geq l$. A constant function is lower semicontinuous. Therefore, $\bar{F} \geq l$ and $\inf \bar{F} \geq \inf F$. Implying that the infimum of both F and its regularization $\bar{F}$ coincide, i.e. $\inf \bar{F} = \inf F$.

Theorem 1.7 (Maxima and Minima)**.** *Let X be a compact metric space and $f : X \to \mathbb{R}$ is continuous, real-valued function. Then f is bounded on X and attains its maximum and minimum. That is, there are x, y belonging to X such that,*

$$f(x) = \inf_{z\in X} f(z) \qquad \text{and} \qquad f(y) = \sup_{z\in X} f(z)$$

Continuity is a strong requirement. Luckily, we can assure the existence of a minimizer of lower-semicontinuous functionals (or maximizer for upper-semicontinuity). The usual procedure to prove existence of a minimizer is making use of Weierstrass' criterion. We take a minimizing sequence and then we prove that the space in which we are trying to find a minimizer element is compact.

Theorem 1.8 (Weierstrass' criterion for existence of minimizers). *If $f : X \to \overline{\mathbb{R}}$ is lower semi-continuous and X is compact, then there exists $\hat{x} \in X$ such $f(\hat{x}) = \min\{f(x) : x \in X\}$.*

Proof. Define $l := \inf\{f(x) : x \in X\} \in \overline{\mathbb{R}}$, notice that $l = +\infty$ only if f is identically $+\infty$, then this case is trivial since any point minimizes f. By compactness there exists a minimizing sequence x_n, that is $f(x_n) \to l$. By compactness we can extract a subsequence converging to some $\hat{x}$ such that $\hat{x} \in X$. By lower-semicontinuity of f, we have that $f(\hat{x}) \liminf_n f(x_n) = l$. Since l is the infimum $l \leq f(\hat{x})$. This proves that $l = f(\hat{x}) \in \mathbb{R}$. □

We can apply the above analysis using a notion of upper-semicontinuity and compactness to find the maximum.

Definition 1.27 (Topological Dual). *If X is a normed space, the dual space $X^* = \mathcal{B}(X, \mathbb{R})$. Consists of all linear and bounded functionals mapping from X to $\mathbb{R}$.*

Definition 1.28 (Weak compactness in dual spaces). *A sequence x_n in a Banach space X is said to be weakly converging to x, and we write $x_n \rightharpoonup x$, if for every $\xi \in X^*$. We have $\langle \xi, x_n \rangle \to \langle \xi, x \rangle$. A sequence $\xi_n \in X^*$ is said to be weakly-* converging to ξ, and we write $\xi_n \rightharpoonup \xi$, if for every $x \in X$ we have $\langle \xi_n, x \rangle \to \langle \xi, n \rangle$.*

Theorem 1.9 (Banach-Alaouglu). *If X is separable and ξ_n is a bounded sequence in X^*, then there exists a subsequence ξ_{n_k} weakly converging to some $\xi \in X^*$.*

The Banach-Alaouglu's theorem is a well known result in functional analysis, an equivalent formulation is saying the closed unit ball in X^* is weak-* compact.

Measure Theory

The optimal transport problem theory is based mostly on Measure Theory. We present some abstract objects and theorems needed to develop in proper way the problem. For better understanding on Measure Theory we refer [1].

Definition 1.29 (Sigma Algebra). *An algebra of sets $\mathcal{A}$ is a class of subsets of some fixed set X (called the space) such that,*

- *X and $\emptyset$ belong to $\mathcal{A}$.*
- *If $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$, $A \cup B \in \mathcal{A}$, $A \backslash B \in \mathcal{A}$.*

An algebra of sets is a σ-algebra if for any sequence of sets $A_n \in \mathcal{A}$ we have $\mathcal{A} \ni \bigcup_{n\in\mathbb{N}} A_n$.

Definition 1.30 (Measure Space). *A pair $(X, \mathcal{A})$ consisting of a set X and a σ-algebra $\mathcal{A}$ of its subsets is called a measurable space.*

Definition 1.31 (Borel σ-algebra). *The Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ of $\mathbb{R}^n$ is the σ-algebra generated by all open sets. The sets in a Borel σ-algebra are called Borel sets. For any set $E \subset \mathbb{R}^n$, let $\mathcal{B}(E)$ denote the class of all sets of the form $E \cap B$, where $B \in \mathcal{B}(\mathbb{R}^n)$.*

Given a topological space, in this text consider the Borel σ-algebra unless stated otherwise.

Definition 1.32 (Measure). *A real-valued set function $\mu : \mathcal{A} \to \overline{\mathbb{R}}$ on a class of sets $\mathcal{A}$ is called countably additive if*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

for all pairwise disjoint sets A_n in $\mathcal{A}$ such that $\mathcal{A} \ni \bigcup_{n=1}^{\infty} A_i$. A countably additive set function defined on an algebra is called a measure.

Given a measure space X, we denote $\mathcal{M}(X)$ and $\mathcal{M}_+(X)$ to refer the set of finite measures and positive finite measures on X, respectively.

Definition 1.33. *A countably additive measure μ on a σ-algebra of subsets of a space X is called a probability measure if $\mu \geq 0$ and $\mu(X) = 1$. A triple $(X, \mathcal{A}, \mu)$ is called a measure space if μ is a nonnegative measure on a σ-algebra A of subset of a set X. If μ is a probability measure, then $(X, \mathcal{A}, \mu)$ is called a probability space.*

Definition 1.34 (Lebesgue Measure)**.**

Definition 1.35 (Hausdorff Measure)**.**

Definition 1.36 (Probability)**.** *A countably additive measure μ on a σ-algebra of subsets of a space X is called a probability measure if $\mu \geq 0$ and $\mu(X) = 1$.*

A triple $(X, \mathcal{A}, \mu)$ is called a measure space if μ is a nonnegative measure on a σ-algebra $\mathcal{A}$ of subset of a set X. If μ is a probability measure, then $(X, \mathcal{A}, \mu)$ is called a probability space.

Definition 1.37 (Product σ-algebra)**.** *Let $(X_1, \mathcal{A}_1, \mu_1)$ and $(X_1, \mathcal{A}_1, \mu_1)$ be two spaces with finite non-negative measures. On the space $X_1 \times X_2$ we consider sets of the form $A_1 \times A_2$, where $A_i \in \mathcal{A}_i$, called measurable rectangles. Let $\mu_1 \times \mu_2\,(A_1 \times A_2) := \mu_1(A_1)\mu_2(A_2)$.*

Let $A_1 \otimes A_2$ denote the σ-algebra generated by all measurable rectangles; this σ-algebra is called the product of the σ-algebras A_1 and A_2.

Theorem 1.10. *The set function $\mu_1 \times \mu_2$ is countably additive on the algebra generated by all measurable rectangles and uniquely extends to a countably additive measure, denoted by $\mu_1 \otimes \mu_2$.*

Definition 1.38 (Image Measure)**.** *Let $(X, \mathcal{A}_X)$ and $(Y, \mathcal{A}_Y)$ be two measurable spaces. Let $T : X \to Y$ be a measurable map from X to Y. Let μ be a measure $\mu : \mathcal{A}_X \to \overline{\mathbb{R}}_+$, then the image measure (or pushforward measure) $T_\#\mu : \mathcal{A}_2 \to \overline{\mathbb{R}}_+$ is given by,*

$$T_\#\mu(B) = \mu\,(T^-1(B)), \quad \forall B \in \mathcal{A}_Y.$$

Theorem 1.11. *Let $(X, \mathcal{A}_X)$ and $(Y, \mathcal{A}_Y)$ be two measurable spaces. Let μ be a nonnegative measure. A $\mathcal{A}_2$-measurable function g on Y is integrable with respect the measure $\mu \circ f^{-1}$ precisely when the function $g \circ f$ is integrable with respect to μ. In addition we have,*

$$\int_Y g(y)\mu \circ f^{-1}(\mathrm{dy}) = \int_X g(f(x))\mu(\mathrm{dx})$$

The space of Borel probability measures on X is denoted by $\mathcal{P}(X)$. The weak topology on $\mathcal{P}(X)$ is induced by convergence against bounded continuous test functions on X, that is $C_b(X)$.

Definition 1.39 (Atom and atomless measures)**.** *The set $A \in \mathcal{A}$ is called an atom of the measure μ if $\mu(A) > 0$ and every set $B \subset A$ from $\mathcal{A}$ has measure either 0 or $\mu(A)$. If there are no atoms, then the measure μ is called atomless.*

A measure over a set $\Omega \subset \mathbb{R}$ is atomless if $\forall x \in \Omega$, we have $\mu(\{x\}) = 0$. The Dirac's measure is not atomless.

Definition 1.40 (Absolutely continuity and singularity)**.** *Let μ and ν be countably additive measures on a measurable space $(X, \mathcal{A})$.*

- *The measure ν is called absolutely continuous with respect to μ if $|\nu|\,(A) = 0$ for every set A with $|\mu| = 0$. We use the notation $\nu \ll \mu$.*
- *The measure ν is called singular with respect to μ if there exists a set $A \in \mathcal{A}$ such that*

$$|\mu|\,(A) = 0 \quad \text{and} \quad |\nu|\,(X \backslash A) = 0$$

If $\nu \ll \mu$ and $\mu \ll \nu$, then the measures μ and ν are equivalent. We use the notation $\mu \sim \nu$ to refer this situation.

The above definition allows us to introduce the Radon-Nikodym theorem that is one of the main results in measure theory.

Theorem 1.12 (Radon–Nikodym theorem). *Let μ and ν be two finite measures on a space $(X, \mathcal{A})$. The measure ν is absolutely continuous with respect to the measure μ precisely when there exists a μ-integrable function f such that ν is given by*

$$\nu(A) = \int_A f \mathrm{d}\mu$$

Definition 1.41 (L^p Spaces).

Theorem 1.13 (Lebesgue dominated convergence theorem). *Suppose that μ-integrable functions f_n converge almost everywhere to a function f. If there exists a μ-integrable function Φ such that,*

$$|f_n|(x) \leq \Phi(x), \quad \text{almost everywhere for every } n$$

then the function is integrable and

$$\int_X f(x)\mathrm{d}\mu(x) = \lim_{n\to\infty} \int_X f_n(x)\mathrm{d}\mu(x)$$

In addition,

$$\lim_{n\to\infty} \int \left|f(x) - f_n(x)\right| \mathrm{d}\mu(x) = 0$$

Theorem 1.14 (Monotone Convergence). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of μ-integrable functions such that $f_n(x) \leq f_{n+1}$ almost everywhere for each $n \in \mathbb{N}$. Suppose that*

$$\sup_{n\in\mathbb{N}} \int_X f_n(x)\mathrm{d}\mu(x) < \infty \tag{1.2}$$

Then the function $f(x) = \lim_{n\to\infty} f_n(x)$ is almost everywhere finite and integrable. In addition the following equality holds,

$$\int_X f(x)\mathrm{d}\mu(x) = \lim_{n\to\infty} \int_X f_n(x)\mathrm{d}\mu(x)$$

Theorem 1.15 (Fatou's Theorem). *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative μ-integrable functions convergent to a function f almost everywhere and let*

$$\sup_{n\in\mathbb{N}} \int_X f_n(x)\mathrm{d}\mu(x) \leq K < \infty$$

Then, the function f is μ-integrable and

$$\int_X f(x)\mathrm{d}\mu(x) \leq K$$

Moreover,

$$\int_X f(x)\mathrm{d}\mu(x) \leq \liminf_{n\to\infty} \int_X f_n(x)\mathrm{d}\mu(x)$$

Corollary 1.1. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of nonnegative μ-integrable functions such that*

$$\sup_{n\in\mathbb{N}} \int_X f_n(x)\mathrm{d}\mu(x) \leq K < \infty$$

Then the function $\liminf_{n\to\infty} f_n$ is μ-integrable and one has

$$\int_X \liminf_{n\to\infty} f_n(x)\mathrm{d}\mu(x) \leq \liminf_{n\to\infty} \int_X f_n(x)\mathrm{d}\mu(x) \leq K \tag{1.3}$$

Corollary 1.2. *The dominated convergence theorem and Fatou's theorem remain valid if in place of almost everywhere convergence in their hypotheses we require convergence of* $(f_n)_{n\in\mathbb{N}}$ *to* f *in measure* μ.

Definition 1.42 (Tightness)**.** *Let* $(X, \mathcal{T})$ *a topological space, and let* $\mathcal{A}$ *a* σ*-algebra on* X *that contains the topology* $\mathcal{T}$. *Let* M *be a collection of measures defined on* $\mathcal{A}$. *The collection* M *is called tight if for every* $\epsilon > 0$ *there is a compact subset* K_ϵ *of* X *such that, for all measures* $\mu \in M$ *we have,*

$$|\mu|\,(X \backslash K_\epsilon) < \epsilon$$

Definition 1.43. *A sequence* μ_n *probability measures over* X *is said to be tight if for every* $\epsilon > 0$, *there exists a compact subset* $K \subset X$ *such that* $\mu_n\,(X \backslash K) < \epsilon$ *for every* n.

Theorem 1.16 (Prokhorov)**.** *Suppose that* μ_n *is a tight sequence of probability measures over a Polish space* X. *Then there exists* $\mu \in \mathcal{P}(X)$ *and a subsequence* μ_{n_k} *such that* $\mu_{n_k} \rightharpoonup \mu$, *in duality with* $C_b(X)$. *Conversely, every sequence* $\mu_{n_k} \rightharpoonup \mu$ *is tight.*

Definition 1.44. *Let* $(X, \mathcal{A}, \mu)$ *be a probability space. Then every Borel-measurable mapping* $\mathcal{X} : X \to \mathbb{R}$ *with for all* $B \in \mathcal{B}(\mathbb{R})$ *is a random variable, denoted by* $\mathcal{X} : (X, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Definition 1.45 (Duality between C_0 and $\mathcal{M}$)**.**

Explanation about notions of convergence with bounded functionals and vanishing in infinity functions. If X is compact we have $C_0(X) = C_b(X) = C(X)$ if X and both notions of convergence coincide.

Theorem 1.17 (Rademacher)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a Lipschitz continuous function. Then the set of points where* f *is not differentiable is negligible for the Lebesgue measure.*

Lemma 1.1. *If* μ, ν *are two probability measures on the real line* $\mathbb{R}$ *and* μ *is atomless, then there exists at least a map* T *such that* $T_\#\mu = \nu$.

Lemma 1.2. *There exists a Borel map* $\sigma_d : \mathbb{R}^d \to \mathbb{R}$ *which is injective, its image is a Borel subset of* $\mathbb{R}$, *and its inverse map is Borel measurable as well.*

Theorem 1.18. *If* μ *and* ν *are two probability measures on* $\mathbb{R}^d$ *and* μ *is atomless, then there exists at least a map* T *such that* $T_\#\mu = \nu$.

2

Basics in Convex Analysis.

Definition 2.1 (Convexity)**.** *Let X a linear space.*

Definition 2.2 (Graph and Epigraph)**.**

Definition 2.3 (Infimal Convolution)**.**

Theorem 2.1 (Hahn-Banach separation theorem)**.**

Theorem 2.2 (Hahn-Banach separation thoerem. Geometric version.)**.**

Definition 2.4 (Extreme Point)**.** *A point x in a convex set C is said to be an extreme point of C if there are no two distinct points x_1 and x_2 in C such that $x = \alpha x_1 + (1-\alpha)\, x_2$ for some $0 < \alpha < 1$.*

Definition 2.5 (Convex conjugate function)**.** *Let X be a Banach space, let $f : X \to \overline{\mathbb{R}}$ be a functional over X. We call the convex conjugate to the function $f^* : X^* \to \overline{\mathbb{R}}$, defined as*

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

Proposition 2.1. *The convex conjugate $f^* : X^* \to \overline{\mathbb{R}}$ of a function $f : X \to \overline{\mathbb{R}}$ is convex.*

Proof. Let x^*, y^* elements of the dual space X^*, and $t \in [0,1]$,

$$\begin{aligned}
f^*\left(tx^* + (1-t)y^*\right) &= \sup_{x\in X} \{\langle tx^* + (1-t)y^*, x\rangle - f(x)\} \\
&= \sup_{x\in X} \{\langle tx^* + (1-t)y^*, x\rangle - tf(x) - (1-t)f(x)\} \\
&= \sup_{x\in X} \{t\,\langle x^*, x\rangle + (1-t)\,\langle y^*, x\rangle - tf(x) - (1-t)f(x)\} \\
&\le \sup_{x,y\in X} \{t\,\langle x^*, x\rangle + (1-t)\,\langle y^*, y\rangle - tf(x) - (1-t)f(y)\} \\
&= t\sup_{x\in X} \{\langle x^*, x\rangle - f(x)\} + (1-t)t\sup_{y\in X} \{\langle y^*, y\rangle - f(y)\} \\
&= tf^*(x^*) + (1-t)f^*(y^*).
\end{aligned}$$

Therefore f^* is convex regardless the convexity of f. □

Theorem 2.3. *A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex and lower-semicontinuous if and only if $f^{**} = f$.*

Lemma 2.1 (Convex envelope theorem)**.** *Let X be a reflexive Banach Space. Then the convex conjugate function f^* is the maximum convex functional below f (also called convex envelope), i.e. if ϱ is convex functional and $\varrho(x) \le f(x)$, $\forall x \in X$. Then, $f^{**}(x) \le f(x)$, and $\varrho(x) \le f^{**}(u)$, $\forall x \in U$. In particular $f^{**} = f$ if and only if J is convex.*

Definition 2.6 (Legendre Transform). *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a convex function, we call the Legendre transform f^**

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{x \cdot y - f(x)\}$$

Corollary 2.1. *A function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is convex and l.s.c. if and only if $f^{**} = f$.*

Definition 2.7 (Subdifferential). *Given a proper convex function $f : X \to (-\infty, \infty]$, the subdifferential of such a function is the mapping $\partial f : X \to X^*$ defined by,*

$$\partial f(x) = \{x^* \in X^*; f(x) - f(y) \leq \langle x^*, x - y \rangle,\ \forall y \in X\}$$

Theorem 2.4 (Geometrical version of Hahn-Banach Theorem). *Let M be a vector subspace of the topological vector space X. Suppose K is a non-empty convex open subset of X with $K \cap M = \emptyset$. Then there is a closed hyperplane $N \in X$ containing M with $K \cap N = \emptyset$.*

Theorem 2.5. *The epigraph of a convex and lower semicontinuous function is a closed convex set in $\mathbb{R}^d \times \mathbb{R}$, and can be written as the intersection of the half-spaces which contain it.*

Here we write the proof for the identity for the projection onto an affine set

Definition 2.8 (Projection onto a Set).

Theorem 2.6.

An important

Definition 2.9 (Duality).

3

Linear Programming

Linear programming is a well studied branch of the mathematics that studies the optimization of linear functions under linear constrains. The study of linear programming started during the second part of the 1940s, as a technique military oriented problems.

We can formulate the problem in its general form as follows:

Problem 1. *Given a cost vector* $\mathbf{c} \in \mathbb{R}^n$*, a linear operator* $\mathbf{A} \in M^{m\times n}$*, the problem consists in finding* $\mathbf{x} \in \mathbb{R}^n$ *such that*

$$\min \quad \mathbf{c}^\top \mathbf{x} \tag{3.1}$$

$$\textit{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.2}$$

$$\mathbf{x} \geq 0 \tag{3.3}$$

We refer to this formulation as the ***primal****.*

Where $\mathbf{A}$ is a $m \times n$ matrix, and $\mathbf{b} \in \mathbb{R}^m$ is an m-dimensional column vector. The vector inequality $\mathbf{x} \geq 0$ means that each component is nonnegative. This problem has a solution if $n > m$.

Definition 3.1. *Given the set of* m *simultaneous linear equations* (3.2) *with* n *unknowns, let* $\mathbf{B}$ *be any nonsingular* $m \times m$ *submatrix made up of columns of* $\mathbb{A}$*. Then if all* $n - m$

Definition 3.2. *If one or more of the basic variables in a basic solution has value zero, that solution is said to be degenerate solution basic solution*

Theorem 3.1 (Fundamental theorem of linear programming.)**.** *Given a linear program in the standard form* (3.1)*,* (3.2) *and* (3.3) *where* $\mathbf{A}$ *is a* $m \times n$ *matrix of rank* m*,*

- *if there is a feasible solution, there is a basic feasible solution.*
- *if there is an optimal solution, there is an optimal basic feasible solution.*

Since for a problem having n variables and m constraints there are at most

$$\binom{n}{m} = \frac{n!}{m!\,(n-m)\,!}$$

basic solutions, the fundamental theorem of linear programming simplifies the problem to a finite number of possibilities. This is a powerful theoretical result, but practical represents an inefficient method to find an optimal solution. This result has an interesting connection to convexity

Theorem 3.2. *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix of rank* m *and* $\mathbf{b}$ *an* m*-vector. Let K be the convex polytope consisting of all* n*-vectors* $\mathbf{x}$ *satisfying*

$$\begin{aligned} \mathbf{A}\mathbf{x} &= \mathbf{b} \\ \mathbf{x} &\geq 0 \end{aligned} \tag{3.4}$$

A vector $\mathbf{x}$ *is an extreme point of* K *if and only if* $\mathbf{x}$ *is a basic feasible solution of* (3.4)*.*

Corollary 3.1. *If the convex set K corresponding to* (3.4) *is nonempty, it has at least one extreme point.*

Corollary 3.2. *If there is a finite optimal solution to a linear programming problem, there is a finite optimal solution which is an extreme point of the constraint set.*

Corollary 3.3. *The constraint set K corresponding to* (3.4) *possesses at most a finite number of extreme points.*

Proof. There is only a finite number of basic solutions generated by selecting m basis vectors and n columns of $\mathbf{A}$. The extreme points of K are a subset of the basic solutions. □

Corollary 3.4. *If the convex polytope K corresponding to* (3.4) *is bounded, then K is a convex polyhedron. That is, K consists of points that are convex combinations of a finite number of points.*

Duality

Problem 2. *Given a cost vector $\mathbf{c} \in \mathbb{R}^n$, a linear operator $\mathbf{A} \in M^{m\times n}$ and a column vector. We say that the dual for the primal formulation 1 is given by,*

$$\begin{aligned} &\max && \boldsymbol{\lambda}^\top \mathbf{b} && (3.5)\\ &\textit{subject to} && \boldsymbol{\lambda}^\top \mathbf{A} \leq \mathbf{c} && (3.6)\\ & && \boldsymbol{\lambda} \geq 0 && (3.7)\end{aligned}$$

Lemma 3.1 (Weak Duality lemma)**.** *If $\mathbf{x}$ and $\boldsymbol{\lambda}$ are feasible for* (3.2) *and* (3.6)*, respectively then $\mathbf{c}^\top\mathbf{x} \geq \boldsymbol{\lambda}^\top\mathbf{b}$*

Proof. We see that following inequality holds for equations (3.2), (3.6) and the cone $\mathbf{x} \geq 0$,

$$\boldsymbol{\lambda}^\top \mathbf{b} = \boldsymbol{\lambda}^\top (\mathbf{A}\mathbf{x}) \leq \mathbf{c}^\top \mathbf{x} \tag{3.8}$$

□

Corollary 3.5. *If $\mathbf{x}_0$ and $\boldsymbol{\lambda}_0$ are feasible for the* (3.2) *and* (3.6) *respectively and $\mathbf{c}^\top\mathbf{x}_0 = \boldsymbol{\lambda}_0^\top\mathbf{b}$, then $\mathbf{x}_0$ and $\boldsymbol{\lambda}_0$ are optimal for their respective problems.*

This corollary is the result of the Weak Duality lemma. Since a feasible vector to the primal problem yields an upper bound on the value of the dual problem. In the other hand, a feasible vector to the dual problem yields a lower bound on the value of the primal problem. The values associated with the primal problem are all larger than the values associated with the dual problem. We see that having a feasible pair $\mathbf{x}_0$ and $\boldsymbol{\lambda}_0$ for their respective problems, satisfying the equality means that each problem has reached its optimal value.

Theorem 3.3 (Duality Theorem)**.** *If the problem* (1) *has a finite optimal solution then the dual formulation* (2) *also does. In the same manner, if the dual problem* (2) *has solution then the primal also does. Moreover the corresponding values of the objective functions are equal. If either problem has an unbounded objective solution, the other problem has no feasible solution.*

Proof. We see from corollary 3.5 that the first condition holds. □

Complementary Slackness.

Simplex Method.

Simplex method is not a polynomial-time algorithm.

Interior Methods.

4

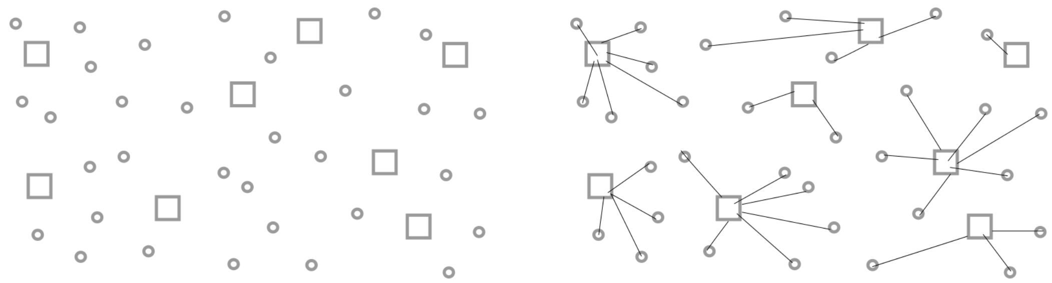
Optimal Transport Theory

To introduce the optimal transport problem please imagine we are asked by a consortium of factories to design a plan for distributing their products among its many customers in such a way that the transportation costs are minimal.

We can start the approach of this problem considering the customers as members of the set X and the factories as members of a set Y. We want to know which factory $y \in Y$ is going to supply a customer $x \in X$, i.e. we represent such assignation of a factory to a customer as map $y = T(x) \in Y$. Therefore, we can estimate the transportation cost $c(x, T(x))$ of supplying a customer x with a factory $y = T(x)$.

We see that our problem is reduced to find an assigning map from the set of customers to the set of factories in such a way that the total cost $C(X, Y) = \sum_{x \in X} c(x, T(x))$ is minimal.

Figure 4.1: Illustration of the problem of Factories supplying customers.



(a) Factories represented by squares. customers represented by circles.

(b) Factories represented by squares. customers represented by circles. Assignation of a factory to a customer represented by a line.

Gaspard Monge was a French mathematician who introduced for the very first time the optimal transport problem as *déblais et remblais* in 1781. Monge was interested in finding a map that distributes an amount of sand or soil extracted from the earth or a mine distributed according to a density f, onto a new construction whose density of mass is characterized by a density g, in such a way the average displacement is minimal. We see that Monge presented a more continuous flavor of the problem.

We remark that we are not interested in the quantity of mass we are transporting. This information it is not relevant for the problem or has no sense its consideration (for example the factories-customer problem). We are interested in finding a way to assign or distribute elements among two sets. We are interested in applications concerning the transportation of a finite amount of mass. Therefore, it is reasonable to state our problem in terms of probability measures.

Formally, given two densities of mass f and g, Monge was interested in finding a map $T : \mathbb{R}^3 \to \mathbb{R}^3$ pushing the one onto the other,

$$\int_A g(y)\mathrm{d}y = \int_{T^{-1}(A)} f(x)\mathrm{d}x$$

For any Borel subset $A \subset \mathbb{R}^3$. And the transport also should minimize the quantity,

$$\int_{\mathbb{R}^3} |x - T(x)|\, f(x)\mathrm{d}x$$

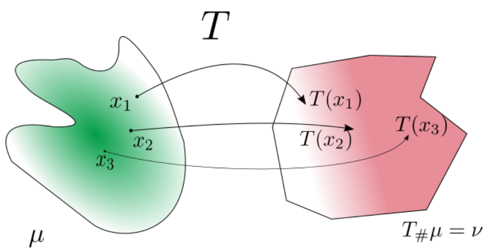
Therefore, we need to search for the optimum in the set of measurables maps $T : X \to Y$ such that the condition (4) is translated to,

$$(T_{\#}\mu)(A) = \mu(T^{-1}(A)) \qquad \text{for every measurable set } A \subset X. \tag{4.1}$$

In other words, we need $T_{\#}\mu = \nu$. Notice that given the context for which the problem was formulated, originally it was binded to $\mathbb{R}^3$ or $\mathbb{R}^2$ but we can consider the general case in $\mathbb{R}^d$. In the Euclidean frameworks if we assume f, g and T regular enough and T also injective, this equality implies,

$$g(T(x)) \det(\mathrm{D}T(x)) = f(x) \tag{4.2}$$

Figure 4.2: Monge problem. Finding a map.



The equation (4.2) is nonlinear in T making difficult the analysis of the Monge's Problem. Moreover, the constrain makes this problem hard to handle since it is not close even under weak convergence.

To appreciate this fact, consider $\mu = \mathscr{L}^1 \llcorner [0,1]$ and the hat functions h_k defined as follow,

$$h_k(x) = \begin{cases} 2kx & x \in \left[0, \frac{1}{2k}\right] \\ 2 - 2kx & x \in \left(\frac{1}{2k}, \frac{1}{k}\right] \\ 0 & \text{otherwise} \end{cases}$$

Then take the sequence $f_n : [0,1] \to [0,1]$,

$$f_n(x) = \sum_{i=0}^{n-1} h_n\left(x - \frac{i}{n}\right) \tag{4.3}$$

We see that the sequence satisfies $f_{n\#}\mu = \mu$. It is easy to check that $\mu\left(f_n^{-1}(A)\right) = \mathscr{L}^1(A)$ for every open set $A \in [0,1]$. In the other hand, the sequence converges weakly to $f_n \rightharpoonup f = \frac{1}{2}$, which obviously makes $f_{\#}\mu \neq \mathscr{L}^1 \llcorner [0,1]$.

Problem 3. *Given two probability measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a cost function $c : X \times Y \to \{0, +\infty\}$, the Monge's problem consists in finding a map $T : X \to Y$*

$$\inf\left\{ M(T) := \int_X c(x, T(x))\mathrm{d}\mu(x) : \quad T_{\#}\mu = \nu \right\} \tag{MP}$$

Monge analyzed geometric properties of the solution to this problem. Although, the question of the existence of a optimal map stayed open until a Russian mathematician named Leonid Vitaliyevich Kantorovich introduced in the paper [2] a suitable framework to study its optimality conditions and prove existence of a minimizer.

When we formulate our factories-customer problem through finding an assignation map, we are excluding the situations in which one customer can be supplied by two or more factories, or in the case of the Monge's problem we are ignoring the possibility of splitting a unit of mass into small pieces that can be assigned simultaneously to different places.

The idea behind Kantorovich's formulation is to consider the transportation maps from one space to another as transportation plans, that is joint probability measures with their marginals given by the initial and final configurations.

Instead of assigning an element of Y to each element of the set X, we can see the problem from a different perspective and assign a weight to the importance of the point $(x, y) \in X \times Y$. We would like to know how much of our total material is distributed from x to y, in such a way to be consistent with information we have the initial and final material configuration. That is, we would like to know the optimal way to concentrate mass to the points (x, y) in such a way we are not creating neither destroying mass.

Designing the transportation strategy using the above procedure is called a transport plan. In terms of probability theory, we are constructing a joint probability measure for $X \times Y$ with marginals given by the measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$.

Please note that in contrast to a map, we can always assign to a point $x \in Y$ as many points in Y as we want, just considering the constraints given by the densities μ and ν. We introduce the following notation to give the necessary formalism to this approach.

Definition 4.1 (Coupling). *Let μ and ν be probability measures of a probability space $(X, \mathcal{A}_X)$ and $(Y, \mathcal{A}_Y)$. Finding a coupling between μ and ν means to construct a measure γ on the space $X \times Y$ (precisely on the product σ-algebra $\mathcal{A}_X \otimes \mathcal{A}_Y$) such that μ and ν are admitted as marginals on X and Y respectively. That is $\mathrm{proj}_{x_\#}\gamma = \mu$ and $\mathrm{proj}_{y_\#}\gamma = \nu$.*

The above definition is equivalently to say that coupling two measures means to find a probability measure γ, such that for all measurable sets $A \subset X$ and $B \subset Y$, one has $\gamma[A \times Y] = \mu[A]$, $\gamma[A \times X] = \nu[B]$.

Moreover, for all integrable (nonnegative measurable) functions ϕ, ψ on X and Y,

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

Since definition 4.1 is given for measures on probabilistic spaces, we can rephrase it in terms of stochastic variables. Let (X, μ) and (Y, ν) be two probability spaces. Coupling μ and ν means constructing two random variables $\mathcal{X}$ and $\mathcal{Y}$ on some probability space, such that $\mathrm{law}(\mathcal{X}) = \mu$, $\mathrm{law}(\mathcal{Y}) = \nu$. The couple $(\mathcal{X}, \mathcal{Y})$ is called a coupling of (μ, ν).

Notice that this approach to solve the problem is more general, since we can always create a transportation plan given a transportation map, i.e.

$$(\mathrm{id}, T)_\# \mu = \gamma \in \mathcal{P}(X \times Y)$$

If T is a transportation map it is easy to check that indeed $\left(\mathrm{proj}_x\right)_\# \gamma = \mu$ and $\left(\mathrm{proj}_y\right)_\# \gamma = \nu$. This inspires a definition for a coupling between two measures generated by a transport map.

Definition 4.2 (Deterministic Coupling). *Let (X, μ) and (Y, ν) be two probabilistic spaces. If there exists a measurable map $T : X \to Y$ such that $T_\# \mu = \nu$. We call the measure $(\mathrm{id}, T)_\# \mu = \gamma \in \mathcal{P}(X \times Y)$ a deterministic coupling of μ and ν.*

For the sake of simplicity, we refer as γ_T a transportation plan generated from a transportation map T.

In terms of stochastic variables, a coupling $(\mathcal{X}, \mathcal{Y})$ is said to be deterministic if there exists a measurable function $T : X \to Y$ such that $\mathcal{Y} = T(\mathcal{X})$. Equivalently, $(\mathcal{X}, \mathcal{Y})$ is a deterministic coupling of μ and ν, if its law $\gamma = \mathrm{law}((\mathcal{X}, \mathcal{Y}))$ is concentrated on the graph of a measurable map $T : X \to Y$. Other way to rephrase it is saying that $\mu = \mathrm{law}(\mathcal{X})$, $\mathcal{Y} = T(\mathcal{X})$, where T is a change of variables from μ to ν, for all ν-integrable (nonnegative measurable) function ϕ,

$$\int_Y \phi(y)\mathrm{d}\nu(y) = \int_X \phi(T(x))\mathrm{d}\mu(x).$$

We use the notation $\Pi(\mu, \nu)$ to refer the **set of couplings** of μ and ν. That is,

$$\Pi(\mu, \nu) = \left\{\gamma \in \mathcal{P}\,(X \times Y) : \left(\mathrm{proj}_x\right)_\# \gamma = \mu \text{ and } \left(\mathrm{proj}_x\right)_\# \gamma = \nu\right\} \tag{4.4}$$

The increasing rearrangement on $\mathbb{R}$ is an example of a coupling between two probability measures over one dimensional euclidean space. Let μ, ν be two probability measures on $\mathbb{R}$. Define their cumulative distribution functions by,

$$F(x) = \int_{-\infty}^{x} \mathrm{d}\mu, \qquad G(y) = \int_{-\infty}^{y} \mathrm{d}\nu$$

Cumulative distributions not always are invertible, since they are not always strictly increasing. Although we can define their pseudo-inverses as follow,

$$F^{-1}(t) = \inf\{x \in \mathbb{R}; F(x) > t\}, \tag{4.5}$$
$$G^{-1}(t) = \inf\{y \in \mathbb{R}; G(y) > t\}. \tag{4.6}$$

Then, we set the map T as $T = G^{-1} \circ F$. If μ is atomless then $T_\# \mu = \nu$.

The increasing rearrangement coupling is useful to construct the *Knothe-Rosenblatt coupling* between two Stochastic variables $\mathbb{R}^n$. Let μ and ν be two probability measures on $\mathbb{R}^n$, such that μ is absolutely continuous with respect to Lebesgue measure. This coupling is constructed in the following way:

1. Take the marginal of the first projection on the first variable; this gives probability measures $\mu_1\,(\mathrm{dx}_1)$, $\nu_1\left(\mathrm{dy}_1\right)$ on $\mathbb{R}$, with μ_1 being atomless. Then define $y_1 = T_1(x_1)$ by the composition of the pseudo-inverse functions of the increasing rearrangement, with F and G considered as they are in (4.5) and (4.6) respectively.

2. Now take the marginal on the first two variables and disintegrate it with respect to the first variable. This gives probability measures $\mu_2\,(\mathrm{dx}_1\mathrm{dx}_2) = \mu_1(\mathrm{dx}_1)\mu_2\,(\mathrm{dx}_2|x_1)$, $\nu_2(\mathrm{dy}_1\mathrm{dy}_2) = \nu_1(\mathrm{dy}_1)\nu_2\left(\mathrm{dy}_2|y_1\right)$. For each given $y_1 \in \mathbb{R}$, we set $y_1 = T_1(x_1)$, and then we define $y_2 = T_2(x_2; x_1)$ under the increasing rearrangement formula of $\mu\,(\mathrm{dx}_2|x_1)$ into $\nu\left(\mathrm{dy}_2|y_1\right)$.

3. We repeat the construction, adding one variable after another. For example, after the assignation $x_1 \to y_1$ has been determined, the conditional probability of x_2 is seen as a one-dimensional probability on a small slice of width dx_1, and it can be transported to the conditional probability of y_2 seen as one dimensional probability of a slice of width dy_1. After n constructions, this procedure maps $\mathcal{Y} = T(\mathcal{X})$.

The *Knothe-Rosenblatt coupling* has the property that its Jacobian matrix of the change of variable T is upper triangular with positive entries on the diagonal.

Lemma 4.1 (Gluing lemma)**.** *If $\mathcal{Z}$ is a function of $\mathcal{Y}$ and $\mathcal{Y}$ is a function of $\mathcal{X}$, then $\mathcal{Z}$ is a function of $\mathcal{X}$. Let (X_i, μ_i), $i = 1, 2, 3$, be Polish probability spaces. If $(\mathcal{X}_1, \mathcal{X}_2)$ is a coupling of (μ_1, μ_2) and $(\mathcal{Y}_2, \mathcal{Y}_3)$ is a coupling of (μ_2, μ_3), then it is possible to construct a triple of random variables $(\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3)$ such that $(\mathcal{Z}_1, \mathcal{Z}_2)$ has the same law as $(\mathcal{X}_1, \mathcal{X}_2)$ and $(\mathcal{Z}_2, \mathcal{Z}_3)$ has the same law as $(\mathcal{Y}_2, \mathcal{Y}_3)$.*

Problem 4. *Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X \times Y \to [0, +\infty]$, we consider the problem*

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

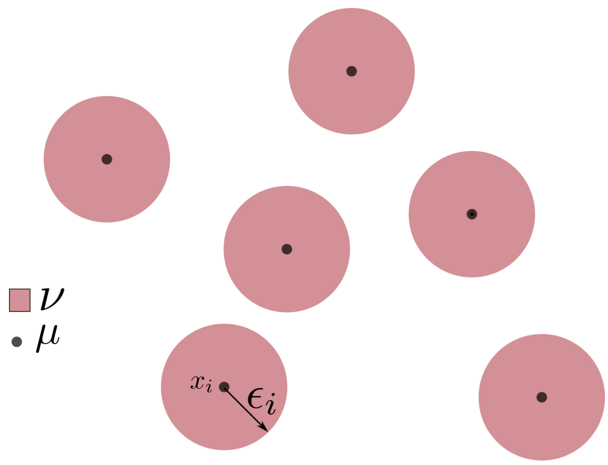
where $\Pi(\mu,\nu)$ is the set of transport plans.

It is a fact for The Kantorovich's formulation that it is always possible to find a transport plan, to see this fact it is enough to take $\gamma = \mu \otimes \nu$. Such a thing it is not always possible with transportation maps (deterministic couplings). For example, consider a measure μ on $\mathbb{R}^d$, concentrated on N different atoms $x_i \in \mathbb{R}^d$,

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} \delta_{x_i}$$

Where δ_{x_i} is the Dirac mass at point x_i. Consider N open balls on $\mathbb{R}^d$ centered at x_i with radius $\epsilon_i > 0$, such that they disjoint pairwise. Let $D = \bigcup_{i=0}^{N-1} B(x_i, \epsilon_i)$ be the union of these balls. Let ν be a the Hausdorff measure of over $D \subset \mathbb{R}^d$. That is $\nu = \mathscr{H} \llcorner D$. We see that it is impossible to couple μ and ν deterministically; since there is no map T, such that $T_{\#}\mu = \nu$.

Figure 4.3: Transportation maps. There is no deterministic coupling for μ and ν, but there is a transportation plan.



Existence of a minimizer for Kantorovich's Problem.

The beauty of Kantorovich's formulation lies on the fact that the set of transport plans is compact under weak convergence making it a suitable framework to use the Weierstrass' criterion to show the existence of a minimizer.

Theorem 4.1. *Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and a cost function $c : X \times Y \to \mathbb{R}$ a continuous function. Then* (KP) *admits a solution.*

Proof. To prove the existence we make use of the Weierstrass' criterion for existence of minimizers. Therefore, we need to prove that $K(\gamma)$ is at least lower semicontinuous and compactness of the space $\Pi(\mu,\nu)$ under some topology.

We choose as a notion of convergence the weak convergence of probability measures in duality with $C_b(X \times Y)$. This immediately implies continuity for $K(\gamma)$ by definition since c is already in $C(X \times Y)$.

Now take a sequence $(\gamma_n)_{n\in\mathbb{N}} \in \Pi(\mu,\nu)$. Since they are probability measures for all n they are bounded in the dual of $C(X \times Y)$. The Banach-Alaouglu's theorem for weak-* compactness in dual spaces guarantees the existence of a convergent subsequence $\gamma_{n_k} \rightharpoonup \gamma$. Let us fix $\phi \in C(X)$ and using $\int \phi(x)\mathrm{d}\gamma_{n_k} = \int \phi \mathrm{d}\mu$ and taking the limit we have $\int_{X\times Y} \phi(x)\mathrm{d}\gamma = \int_X \phi \mathrm{d}\mu$. Proving that $\gamma_{\#}\left(\mathrm{proj}_x\right) = \mu$. We can repeat this argument for ν, fixing $\psi \in C(Y)$ and taking the limit of $\int_{X\times Y} \psi(y)\mathrm{d}\gamma_{n_k} = \int \psi \mathrm{d}\nu$, implies $\int_{X\times Y}\psi(y)\mathrm{d}\gamma = \int \psi \mathrm{d}\nu$. This proves that $\gamma_\#\left(\mathrm{proj}_y\right) = \nu$. Hence, the limit $\gamma \in \Pi(\mu,\nu)$ showing that the set of couplings of μ and ν is sequentially compact. □

Continuity for the cost function and compactness of the metric spaces can be demanding requirements. However we can substitute them by milder conditions for the existence of a minimizer.

Lemma 4.2. *Let X be a metric space. If $f : X \to \overline{\mathbb{R}}$ is a lower semi-continuous function, bounded from below, then the functional $J : \mathcal{M}_+(X) \to \overline{\mathbb{R}}$ defined on the space of finite positive measures on X, given by*

$$J(\mu) = \int f \mathrm{d}\mu$$

is lower semi-continuous for the weak convergence of measures.

Proof. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of continuous and bounded functions, converging increasingly to f. Consider the functionals $J_n : \mathcal{M}_+(X) \to \overline{\mathbb{R}}$, defined as

$$J_n(\mu) = \int f_n \mathrm{d}\mu$$

Every J_n is continuous for the weak convergence. We set $J(\mu) = \int f \mathrm{d}\mu$. We see that $J_n(\mu) \le J(\mu)$ for any μ. Since our functions are bounded, and f is bounded from below, and our measures are finite, we can make use of monotone convergence theorem, $J_n(\mu) \to J(\mu)$, having as result $J(\mu) = \sup_n J_n(\mu)$. Since we have that $J(\mu)$ is the supremum of continuous functions we can assure that that J is lower semicontinuous. □

Theorem 4.2. *Let X and Y be compact metric spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X \times Y \to \overline{\mathbb{R}}$ be lower semi-continuous and bounded from below. Then Kantorovich's problem admits a solution.*

Proof. We make use of lemma 4.2, setting $f = c$ on the space $X \times Y$. We apply again Weierstrass criterion proving existence of a minimizer. □

Theorem 4.3. *Let X and Y be Polish spaces, and $c : X \times Y \to \overline{\mathbb{R}}_+$, a real valued lower semi-continuous cost function on the space $X \times Y$. Then* (KP) *admits a solution.*

Lemma 4.3. *Let X and Y be Polish spaces, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c : X \times Y \to [0, +\infty]$ lower semicontinuous. Then the Kantorovich's problem admits a solution.*

Proof. Fix $\epsilon > 0$ and find two compact sets $K_x \subset X$ and $K_y \subset Y$ such that $\mu\left(X\backslash K_x\right) < \epsilon$, and $\nu\left(Y\backslash K_y\right) < \epsilon$. Then the set $K_X \times K_Y$ is compact in $X \times Y$ and, for any $\gamma_n \in \Pi(\mu,\nu)$, we have,

$$\begin{aligned}
\gamma_n\left((X\times Y)\backslash\left(K_X\times K_Y\right)\right) &\le \gamma_n\left(\left(X\backslash K_X\right)\times Y\right) + \gamma_n\left(X\times\left(Y\backslash K_y\right)\right) && (4.7)\\
&= \mu\left(X\backslash K_X\right) + \nu\left(Y\backslash K_Y\right) && (4.8)\\
&= 2\epsilon && (4.9)
\end{aligned}$$

Given the arbitrary way to choose ϵ, this shows tightness of all sequences in $\Pi(\mu,\nu)$ and hence compactness. □

Kantorovich formulation as relaxation

There are situations in which is possible to find a deterministic coupling between two measures, but not an optimal one for a cost function $c : X \times Y \to \overline{\mathbb{R}}$. A common example, popular in the literature, is the following: consider as cost function $c : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, the Euclidean distance $c(x,y) = |x - y|$, the measure $\mu = \mathscr{H}\llcorner D$ as the Hausdorff measure for the segment $D = \left\{(0,t)^\top \in \mathbb{R}^2 : \text{ for } t \in [0,1]\right\}$.

Let D_1 and D_2 be the segments given by,

$$\begin{aligned}
D_1 &= \left\{(-1,t)^\top \in \mathbb{R}^2 : \text{ for } t \in [0,1]\right\}\\
D_2 &= \left\{(+1,t)^\top \in \mathbb{R}^2 : \text{ for } t \in [0,1]\right\}
\end{aligned}$$

And we set the measure ν as follows,

$$\nu = \frac{\mathscr{H} \llcorner D_1 + \mathscr{H} \llcorner D_2}{2}$$

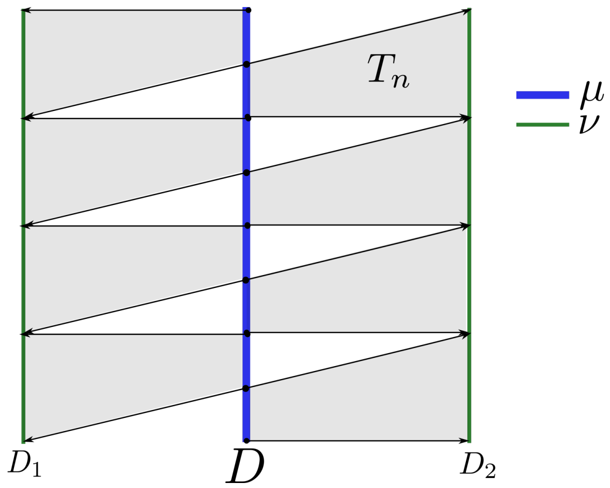


Figure 4.4: There is a deterministic coupling for μ and ν, but no optimal one. The map T_n shown in this picture with $n = 4$.

There are many ways to construct a transportation map for this situation. Consider the maps T_n constructed splitting the segment D into $2n$ equal parts and the segments D_1 and D_2 in n equal parts. We label the parts of the segment D with the integer numbers from 0 to $2n-1$. Then the map T_n assign the parts of D labeled with even numbers to the right hand side segment D_2 and the parts labeled with odd numbers to the left right side segment D_1.

Formally, let $k = 0, \dots, 2n-1$ be an integer used to label the equal parts of D,

$$T_n\left(\begin{pmatrix}0\\t\end{pmatrix} \in D\right) = \begin{cases} \begin{pmatrix}1\\2t - \frac{k}{2n}\end{pmatrix} & k \text{ even and } t \in \left[\frac{k}{2n}, \frac{k+1}{2n}\right), \\ \begin{pmatrix}-1\\2t - \frac{k+1}{2n}\end{pmatrix} & k \text{ odd and } t \in \left(\frac{k}{2n}, \frac{k+1}{2n}\right]. \end{cases}$$

We can find an upper boundary for the total cost $C(T_n)$,

$$\begin{aligned} C(T_n) &= \int_D \left|x - T_n(x)\right| \mathrm{d}\mu(x) \\ &= 2n \int_0^{\frac{1}{2n}} \sqrt{1 + 4t^2} \mathrm{dt} \\ &\leq 2n \left(\int_0^{\frac{1}{2n}} 1 + 4t^2 \mathrm{dt}\right)^{1/2} \left(\int_0^{\frac{1}{2n}} \mathrm{dt}\right)^{1/2} \\ &= \sqrt{1 + \frac{1}{3n^3}} \\ &\leq 1 + \frac{1}{n} \end{aligned}$$

Let γ_{T_n} be deterministic coupling generated by T_n. We see that we can find always find a cheaper plan $\gamma_{T_{n+1}}$ for any $n \in \mathbb{N}$. This sequence of transportation plans converges weakly to the plan $\gamma_{T_n} \rightharpoonup \gamma_T = \frac{\gamma_{T^+}}{2} + \frac{\gamma_{T^-}}{2}$. Where T^+ and T^- are given by:

$$T^+(x) = x + \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$T^-(x) = x - \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

The idea is that the mass of each point $x \in D$ is split in two and equally distributed among D_1 and D_2 assigning one half of the mass respectively. Note that this distribution is an optimal plan for the cost function $c(x, y) = |x - y|$. Because of the triangle inequality, sending the mass from $x \in D$ to any other point of D_1 and D_2 different than those assigned by the maps $T^{\pm}$, implies a higher cost.

From the last example we see that a sequence of deterministic couplings converges to a transportation plan that is a solution for Kantorovich's problem (KP), but clearly it is not for Monge's problem (MP). We also gave one example where (MP) has no solution. Assume for a moment that Monge's situation where indeed does exist a solution for Monge's problem, then the following question arises: Is there any situation where Monge's problem and Kantorovich's problem have the same solution?

Lemma 4.4. *On a compact subset $\Omega \subset \mathbb{R}^d$, the set of plans γ_T induced by a transport is dense in the set of plans $\Pi(\mu, \nu)$ whenever μ is atomless.*

Theorem 4.4. *On a compact subset $\Omega \subset \mathbb{R}^d$, $K\,(\gamma)$ is the relaxation of $J(\gamma)$. In particular, $\inf J = \min K$, and hence Monge and Kantorovich problems have the same infimum.*

Proof. Since K is continuous, then it is lower semicontinuous, and since we have $K \leq J$, then K is necessarily smaller than the relaxation of J. We only need to prove that, for each γ, we can find a sequence of transports T_n such that $\gamma_{T_n} \to \gamma$ and $J\left(\gamma_{T_n}\right) \to K(\gamma)$, so that the infimum in the sequential characterization of the relaxed functional will be smaller than K, thus providing the equality.

Actually, since for $\gamma = \gamma_{T_n}$ be two functionals K and J coincide, and since K is continuous we only need to produce a sequence T_n such that $\gamma_{T_n} \rightharpoonup \gamma$. The last step is possible because of the density of transport plans generated by a map γ_{T^n} in the set of transport plans $\Pi(\mu, \nu)$. ☐

Cyclical Monotonicity and Duality.

Consider a similar situation to the factories-customers example, but in this new hypothetical situation the consortium has already a fixed transportation plan. They know that the costs are high and they want to make them cheaper.

Definition 4.3. *Let X, Y be arbitrary sets, and $c : X \times Y \to (-\infty, \infty]$ be a cost function. A subset $\Gamma \subset X \times Y$ is said to be c-cyclically monotone if, for any $N \in \mathbb{N}$, and any family of points $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$ of Γ, the inequality*

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{i+1})$$

considering $N + 1 = 1$.

Since any permutation σ over the set $\{1, \ldots, N\}$ can be written as a product of disjoint cycles, we have that this property satisfies,

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}) \tag{4.10}$$

The subdifferentials of convex functions on $\mathbb{R}^n$ are characterized in terms of a monotonicity property. **Restate Rockafaller**

Theorem 4.5 (Rockafellar)**.** *Let Γ be a c-cyclically monotone set. In order that there exists a closed proper convex function f on $\mathbb{R}^n$ such that $\Gamma \subset \partial f(x)$ for every x, it is necessary and sufficient that Γ be cyclically monotone.*

The theorem 4.5 is a well known theorem in convex analysis. It basically states that every cyclically monotone set is contained in the graph of the subdifferential of a convex function.

Definition 4.4 (Support of transport plan.). *Given a separable metric space X, the support of a measure γ is defined as the smallest closed set on which γ is concentrated,*

$$\operatorname{spt}(\gamma) := \bigcap_{\substack{\gamma(X\backslash A)=0\\ A=\bar{A}}} A \tag{4.11}$$

Theorem 4.6. *If γ is an optimal transport plan for the cost c and c is continuous, then* spt *is a c-CM set.*

Duality

Imagine instead that the consortium changed its policy and it has decided not to be responsible any longer for the transportation of the goods, letting the customers to solve this problem by themselves (assume that the consortium has the monopoly of the goods and the customers have no choice but to adhere to this policy). An entrepreneur feeling that he can ship the goods more efficiently that the consortium did, he intend to buy the goods at the factories and selling them at the customers' stores. Then, he must negotiate with the consortium the prices $-\phi(x)$ that he is able to pay at each factory for the goods and the selling prices $\psi(y)$ at each customers' store. In order to succeed, he need to be competitive and should do it better than the consortium did. Therefore, he must be able that cover with the difference of the sale prices the transportation costs and they should be less than the consortium's costs $\psi(y) + \phi(x) \leq c(x, y)$. He is subject to this constraint and he should negotiate with the consortium and the customers the prices $\phi(x)$ and $\psi(y)$ in order to obtain the maximum profit.

Note that if $\pi \in \mathcal{M}_+(X \times Y)$, we can $\pi \in \Pi(\mu, \nu)$

$$\sup_{\phi,\psi}\left(\int_X \phi \mathrm{d}\mu + \int_Y \psi \mathrm{d}\nu - \int_{X\times Y} (\phi(x) + \psi(y))\, \mathrm{d}\gamma\right) = \begin{cases} 0 & \text{if } \gamma \in \Pi(\mu,\nu) \\ +\infty & \text{otherwise.} \end{cases} \tag{4.12}$$

Definition 4.5. *Given a function $\omega : X \to \overline{\mathbb{R}}$, we define its c-transform, also referred as $c-$conjugate of ω, the function $\omega^c : Y \to \overline{\mathbb{R}}$ by*

$$\omega^c(y) = \inf_{x\in X} c(x, y) - \omega(x) \tag{4.13}$$

In similar way, we can define the $\bar{c}-$transform of $\upsilon : Y \to \overline{\mathbb{R}}$ by

$$\upsilon^c(x) = \inf_{y\in Y} c(x, y) - \upsilon(y) \tag{4.14}$$

Properties of Optimal plans.

Theorem 4.7 (Convexity of optimal plans). *The set of solutions $\bar{\gamma} \in \Pi(\mu, \nu)$ for the Kantorovich's problem is a convex set.*

Proof. We see immediately that if γ_1 and γ_2 solve the Kantorovich's problem, for any $t \in [0, 1]$, the plan $\gamma = t\gamma_1 + (1 - t)\gamma_2$, also solves the problem. □

aca usamos las definiciones de c-convexity y L2 norm para encontrar el optimal map de una gaussiana.

Wasserstein Spaces. $\mathbb{W}_p$

5

Computation of an Optimal Transport

The approximation of an optimal transport is a challenging problem, computationally speaking. We have found a rich literature on it, and many recent advances in this topic have arisen in the very last years.

Linear Programming Formulation.

Simplex Method Algorithm and Duality.

$$\arg\min_{P \in \Pi(\mu,\nu)} \langle C, P \rangle \tag{5.1}$$

Sinkhorn-Knopp Algorithm.

$$\arg\min_{P \in \Pi(\mu,\nu)} \langle C, P \rangle + \mathbf{H}(P) \tag{5.2}$$

Heuristic Methods.

Continuous Formulation.

Beckman Problem and Optimal Transport.

Proximal Splitting Algorithms.

6

Applications

Isoperimetric Inequality.
Dynamical Optimal transport.
Approximation of Euler Equations.
Track of a Dynamic.

Bibliography

[1] Vladimir I. Bogachev. *Measure theory*. Springer-Verlag, Berlin Heidelberg, 2007.

[2] L. Kantorovich. On the translocation of masses. *Dokl. Acad. Nauk. USSR 37*, pages 7–8, 1942.

[3] J.R. Munkres. *Topology*. Topology. Prentice-Hall, 2000. ISBN 9780131784499.