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.
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{+\infty\}$. That is $[-\infty, \infty]$
$\mathbb{R}_+$	The set of nonnegative real numbers, that is the interval $[0, \infty)$.
$\bar{\mathbb{R}}_+$	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.
$\mathcal{P}(X)$	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 $\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\rho + \epsilon\chi)\Big\|_{\epsilon=0} = \int \frac{\delta F}{\delta \rho} \mathrm{d}\chi$
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.
$\mathcal{L}^p$	Lebesgue measure on $\mathbb{R}^p$

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 [2] 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. 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.*

Definition 1.5 (Hausdorff space)**.**

Definition 1.6 (Distance and metric space)**.**

Definition 1.7 (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 $x_{n_k} \to x \in K$.*

We can see from the above definition that any continuous function is lower-semicontinuous.

Definition 1.8 (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.*

Functional Analysis.

Definition 1.9 (Banach Space)**.**

Definition 1.10 (Hilbert Space)**.**

Definition 1.11 (Proper convex function)**.** *Let $f : X \to \bar{\mathbb{R}}$, 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 $\forall x \in X$,

$$f(x) > -\infty \tag{1.1}$$

Definition 1.12 (Projection of a Cartesian Product.)**.** *Let $\pi_1 : X \times Y \to X$ be defined by the equation*

$$\pi_1(x, y) = x;$$

Equivalently, let $\pi_2 : X \times Y \to Y$ be defined by,

$$\pi_2(x, y) = y$$

This maps π_1 and π_2 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 1.13 (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 $f : X \to \mathbb{R}$ a functional valued in $\bar{\mathbb{R}}$. 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.

Definition 1.14 (Lower Semicontinuity). *On a complete metric space X, a function $f : X \to \bar{\mathbb{R}}$ 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*

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

Theorem 1.2 (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 1.3 (Weierstrass' criterion for existence of minimizers). *If $f : X \to \bar{\mathbb{R}}$ is lower semi-continuous and X is compact, then there exists $\hat{x} \in X$.*

Proof. Define $l := \inf\{f(x) : x \in X\} \in \bar{\mathbb{R}}$, 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 $l \le f(\hat{x})$. 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. **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.

Definition 1.15 (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 1.16 (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 1.4 (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.

Definition 1.17. *Let $F : X \to \bar{\mathbb{R}}$ be a given functional bounded from below on a metric space X. Let $\mathcal{G}$ be the set of lower semicontinuous functions $G : X \to \bar{\mathbb{R}}$, such that $G \le F$. 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:*

$$\bar{F}(x) = \inf\left\{\liminf_{n\in\mathbb{N}} F(x_n) : x_n \to x\right\}. \tag{1.2}$$

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

Measure Theory

Definition 1.18 (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 1.19 (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 1.20 (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 $B \in \mathcal{B}(\mathbb{R}^n)$.*

In this text consider the Borel σ-algebra unless noted otherwise.

Definition 1.21 (Measure). *A real-valued set function $\mu : \mathcal{A} \to \bar{\mathbb{R}}$ 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 $\ni \bigcup_{n=1}^{\infty}$. A countably additive set function defined on an algebra is called a measure.

Definition 1.22. *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 1.23 (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 1.24 (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 1.5. *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 1.25 (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 $\mu : \mathcal{A}_X \to \bar{\mathbb{R}}_+$, then the image measure (or pushforward measure) $T_\# \mu : \mathcal{A}_2 \to \bar{\mathbb{R}}_+$ is given by,*

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

Theorem 1.6. *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 1.26 (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 1.27 (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 1.7 (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 1.28 (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 1.29. *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 1.8 (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 1.30. *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 $\mathcal{X} : (X, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*

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)**.**

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 $0 < \alpha < 1$.*

Proposition 2.1. *Let $f : X \to \bar{\mathbb{R}}$ 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 2.1. *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 2.5 (Convex conjugate function)**.** *Let X be a Banach space, let $f : X \to \bar{\mathbb{R}}$ be a functional over X. We call the convex conjugate to the function $f^* : X^* \to \bar{\mathbb{R}}$, defined as*

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

Proposition 2.2. *The convex conjugate $f^* : X^* \to \bar{\mathbb{R}}$ of a function $f : X \to \bar{\mathbb{R}}$ is convex.*

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

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

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

Theorem 2.2. *A function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ 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 $\varrho(x) \leq f(x)$, $\forall x \in X$. Then, $f^{**}(x) \leq f(x)$, and $\varrho(x) \leq f^{**}(u)$, $\forall x \in U$. In particular $f^{**} = f$ if and only if J is convex.*

Definition 2.6 (Legendre Transform)**.** *Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ 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 $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ 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 2.3 (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 2.4. *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.*

Definition 2.8 (Projection onto a Set)**.**

Theorem 2.5.

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* $\mathbf{A} \in M^{m\times n}$

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

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

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

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

Consider the system of equalities (3.2), the vector $\mathbf{b}$ belongs to $\mathbb{R}^m$.

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* $\mathbb{A}$ *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 theoretical powerful 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

Lemma 3.1 (Weak Duality lemma)**.** *If $\mathbf{x}$ and $\mathbf{v}$ are feasible for and , respectively, then c T xxxx xT b.*

Simplex Method.

Interior Method.

4

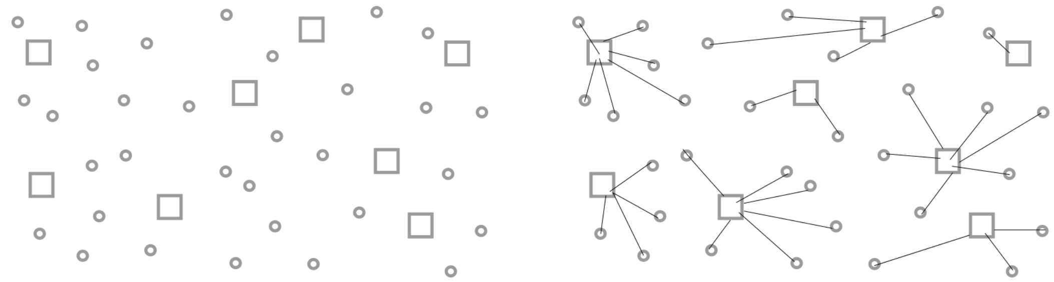
Optimal Transport Theory

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,

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

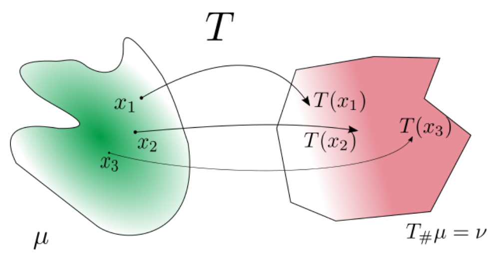
Therefore, we need to search for the optimum in the set of measurables maps $T : X \to Y$ such that the condition (4) is translated to,

$$(T_{\#}\mu)(A) = \mu(T^{-1}(A)) \tag{4.1}$$

for every measurable set A. 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 $\mu = \mathcal{L}^1 \llcorner [0,1]$ 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 $\mu\left(f_n^{-1}(A)\right) = \mathcal{L}^1(A)$ 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 $f_{\#}\mu \neq \mathcal{L}^1 \llcorner [0,1]$.

Problem 2. *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 [1] a suitable framework to study its optimality conditions and prove existence of a minimizer.

Formulating 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 for 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 instead of transportation maps from one space to another, 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 $(x, y) \in X \times Y$. In a better manner we would like to know how much of our total material is distributed in a way (x, y) in such a way to be consistent with information we have the initial and final material configuration. We call this way to design the transportation strategy a transport plan. In terms of probability, 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 that we are not creating or destroying mass.

We introduce the following notation.

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 the measure of the projections π_x and π_y*

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 $(\Omega, \mathcal{P})$ such that $\text{law}(\mathcal{X}) = \mu$, $\text{law}(\mathcal{Y}) = \nu$. The couple $(\mathcal{X}, \mathcal{Y})$ is called a coupling of (μ, ν).

Definition 4.2 (Deterministic Coupling). *A coupling $(\mathcal{X}, \mathcal{Y})$ is said to be deterministic if there exists a measurable function $T : X \to Y$ such that $\mathcal{Y} = T(\mathcal{X})$.*

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,

$$F^{-1}(t) = \inf\{x \in \mathbb{R}; F(x) > t\}; \tag{4.4}$$

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

and we set the map T as,

$$T = G^{-1} \circ F \tag{4.6}$$

We see if μ is atomless then $T_{\#}\mu = \nu$.

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

1. Take the marginal of the first projection on the first variable; this gives probability measures $\mu_1(\mathrm{dx}_1)$, $\nu_1(\mathrm{dy}_1)$ on $\mathbb{R}$, with μ_1 being atomless. Then define $y_1 = T_1(x_1)$ by the formula (4.6) with F and G considered as they are in (4.4) and (4.5) respectively.

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

Lemma 4.1 (Gluing lemma)**.** *Let $(\mathcal{X}_i, \mu_i)$, $i = 1, 2, 3$, be Polish probability spaces. If (X_1, X_2) is a coupling of (μ_1, μ_2) and $(\mathcal{Y}_2, Y_3)$ is a coupling of (μ_2, μ_3), then it is possible to construct a triple of random variables (Z_1, Z_2, Z_3) such that (Z_1, Z_2) has the same law as (X_1, X_2) and (Z_2, Z_3) has the same law as (Y_2, Y_3).*

Notice that this way to see 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 $\pi_{x\#}\gamma = \mu$ and $\pi_{y\#}\gamma = \nu$. The beauty of Kantorovich's formulation lies on the fact that it is always possible to find a transport. Moreover, the space of transport plans is compact.

Problem 3. *Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, and $c : X \times Y \to [0, +\infty]$, we consider the problem*

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

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

Kantorovich formulation as relaxation

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

Cyclical Monotonicity.

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

Definition 4.3 (c-transform)**.** *Let X and Y be sets, and $c : X \times Y \to (-\infty, \infty]$. A function $\psi : X \to \bar{\mathbb{R}}$ is said to be c-convex if it is not identically to $+\infty$ and there exists $\psi^c : Y \to \bar{\mathbb{R}}$*

$$\psi^c(y) = \inf_{x\in X} c(x, y) - \psi(x). \tag{4.7}$$

Definition 4.4. *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 $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$ 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 $\{1, \dots, N\}$ can be written as a product of disjoint cycles, we have that this property satisfies,

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}) \tag{4.8}$$

Definition 4.5 (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,*

$$\mathrm{spt}(\gamma) := \bigcap_{\substack{\gamma(X \setminus A) = 0 \\ A = \bar{A}}} A \tag{4.9}$$

We can fix a point $(x_0, y_0) \in \mathrm{spt}(\gamma)$, then

Theorem 4.2.

Properties of Optimal plans.
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.

Sinkhorn-Knopp Algorithm.

Simulated Annealing.

Continuous Formulation. Beckman Problem and Optimal Transport.

Douglas-Rachford Solver

6

Applications

Isoperimetric Inequality.
Dynamical Optimal transport.
Approximation of Euler Equations.

Bibliography

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

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