Notes on Optimal Transport

Manuel Quintero Coronel

June 25, 2024

1 Introduction to Optimal Transport

Imagine a scenario where a construction company needs to level a piece of land: they need to move soil from several hills (where there's too much soil) to several pits (where there's too little). Each hill and pit has a certain amount of soil to be moved, and transporting soil involves costs—fuel, labor, and wear and tear on machinery—that increase with distance. The company's goal is to move the soil in such a way that the land is leveled at the lowest possible cost. This scenario, simple in its essence, introduces us to the fundamental problem of optimal transport: how to transport mass (in this case, soil) from one configuration to another in the most efficient way possible.

The problem can be abstractly defined as finding the most efficient way to transport mass (modeled as a distribution of matter or probability) from one distribution to another, minimizing the overall transportation cost.

To formalize this scenario let's model the soil from the hills and the soil from the pits by two probability measures, μ, and ν, defined on two measure spaces X and Y, respectively. We also need to pay a price for transporting mass, a **cost function** $c : X \times Y \to [0, +\infty]$ which is measurable and measures the cost of transporting one unit of mass from $x \in X$ to $y \in Y$. Then, the optimal transport problem is how to transport μ to ν whilst minimizing the cost c. We also need to define a way to transfer mass, a **Transference plan**.

Definition 1.1 (Transference Plan). *Let μ and ν be probability measures on metric spaces X and Y respectively. A* transference plan *π is a probability measure on the product space $X \times Y$.*

2 Monge Problem and Kantorovich Relaxation

Definition 2.1 (Pushforward Measure). *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let $T : X \to Y$ be a measurable map. For a measure μ on $(X, \mathcal{A})$, the* pushforward measure *of μ by T, denoted by $T_{\#}\mu$, is a measure on $(Y, \mathcal{B})$ defined for any $B \in \mathcal{B}$ by*

$$T_{\#}\mu(B) := \mu(T^{-1}(B)),$$

where $T^{-1}(B) = \{x \in X : T(x) \in B\}$. *We say* T *transports* μ *onto* ν.

Definition 2.2 (Monge problem). *Consider two metric spaces* X, Y, *two probability measures* $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ *and a cost function* $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$. *Monge's problem is the following optimization problem*

$$\inf_{T:T_{\#}\mu=\nu} \left\{ \int_X c(x, T(x)) d\mu(x) \right\}$$

The map T that attains the infimum is called an **optimal transport plan**. However, there might not exist a map $T : X \to Y$ such that $\nu = \mu \circ T^{-1}$:

Example 2.1. *Let* ν *a probability measure such that* $|Supp(v)| > 1$ *and* $\mu = \delta_a$ *for some* $a \in X$. *Then, for* $B \subset Y$

$$\nu(B) = T_{\#}\mu(B) = \mu(T^{-1}(B)) = \delta_{T(a)}.$$

But clearly this is not possible.

The Kantorovich relaxation extends Monge's problem to a broader class of admissible transport plans, allowing for mass splitting.

Definition 2.3 (admissible transference plan). *An* ***admissible transference plan*** *has* μ *and* ν *as its marginals, meaning that for all measurable sets* $A \subseteq X$ *and* $B \subseteq Y$, *we have:*

$$\pi(A \times Y) = \mu(A) \quad \textit{and} \quad \pi(X \times B) = \nu(B).$$

We denote the set of admissible transference plans by $\Pi(\mu, \nu)$. Note that some equivalent representation of admissibility are

$$\int_Y d\pi(x, y) = d\mu(x) \quad \text{and} \quad \int_X d\pi(x, y) = d\nu(y),$$

and, for all φ, ψ measurable functions, say, in $L^1(\mu) \times L^1(\nu)$,

$$\int_{X \times Y} (\varphi(x) + \psi(y))\, d\pi(x, y) = \int_X \varphi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y).$$

Definition 2.4 (Kantorovich Problem). *Given two probability measures* $\mu \in \mathcal{P}(X)$ *and* $\nu \in \mathcal{P}(Y)$ *on metric spaces* X *and* Y, *and a cost function* $c : X \times Y \to [0, +\infty]$, *the* Kantorovich problem *is defined as:*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi)$$

where

$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(X \times Y) : \pi(A \times Y) = \mu(A) \quad \textit{and} \quad \pi(X \times B) = \nu(B)\}.$$

Then, it is natural to ask if there, and when, exists an optimal transport plan for the Kantorovich problem.

Definition 2.5 (Lower Semicontinuity). *A function $f : X \to \mathbb{R} \cup \{+\infty\}$, where X is a topological space, is said to be* lower semicontinuous *at a point $x \in X$ if for every sequence $(x_n) \subset X$ such that $x_n \xrightarrow[n\to\infty]{} x$, the following condition is satisfied:*

$$\liminf_{n\to\infty} f(x_n) \geq f(x).$$

Remark 2.1. *f is lower semicontinuous $\iff$ epi(f) is closed $\iff$ the sublevel sets $f^{-1}((-\infty, a))$ are closed.*

Proposition 2.1. *Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ where X, Y are Polish spaces, and assume $c : X \times Y \to [0, \infty)$ is lower semicontinuous. Then there exists $\pi^* \in \Pi(\mu, \nu)$ such that*

$$\pi^* \in \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi).$$

The general idea of the proof is to show that the set $\Pi(\mu, \nu)$ is **compact** in the weak topology $\mathcal{P}(X \times Y)$ and that the functional $\mathbb{K}(\pi)$ is lower semicontinuous. Then we would use the general fact that lower semicontinuous functions attain their minima in compact sets. By Prokhorov's Theorem it is sufficient to show that $\Pi(\mu, \nu)$ is closed and uniformly tight.

Proof. Note $\Pi(\mu, \nu) \neq \emptyset$ since $\mu \otimes \nu \in \Pi(\mu, \nu)$.
We first show that $\Pi(\mu, \nu)$ is compact. Let $\delta > 0$, and let $(K, L) \subseteq X \times Y$ be such that $\mu(X \setminus K) \leq \delta$ and $\nu(Y \setminus L) \leq \delta$.

For $\pi \in \Pi(\mu, \nu)$, we have:

$$\pi((X \times Y) \setminus (K \times L)) \underset{(1)}{\leq} \pi(X \times (Y \setminus L)) + \pi((X \setminus K) \times Y)$$

$$= \nu(Y \setminus L) + \mu(X \setminus K) \leq 2\delta.$$

Where (1), follows since $(x, y) \in \pi((X \times Y) \setminus (K \times L)) \implies x \notin K$ or $y \notin L$

$$\Rightarrow (x, y) \in (X \times (Y \setminus L)) \cup ((X \setminus K) \times Y).$$

Thus, $\Pi(\mu, \nu)$ is tight. By Prokhorov's Theorem[1], $\overline{\Pi(\mu, \nu)}$ is sequentially compact (relatively compact with respect to the weak topology).

[1]Prokhorov's Theorem implies that if (π_n) is a tight sequence in $\mathcal{P}(X)$, then there exists a subsequence (π_{n_k}) and a probability measure $\pi \in \mathcal{P}(X)$ such that π_{n_k} converges weakly to π. We are actually using a Corollary of this.

Now, the integral is a continuous operator, which in turn gives that the conditions that define $\Pi(\mu,\nu)$ are continuous with respect to $P(X\times Y)$. Then sequences in $\Pi(\mu,\nu)$ converge to points in $\Pi(\mu,\nu)$ thus containing its limit points, hence closed (under the weak topology). Alternatively, the admissible couplings condition holds for all bounded continuous functions (and by weak convergence it is closed).

Let $(\pi_n)\subseteq\Pi(\mu,\nu)$ be a minimizing sequence (i.e., $K(\pi_n)\to\inf_{\pi\in\Pi(\mu,\nu)}K(\pi)$) such that $\pi_n\to\pi^*$. By compactness of $\Pi(\mu,\nu)$, $\pi^*\in\Pi(\mu,\nu)$. Now, since c is l.s.c. and nonnegative (thus bounded from below), we can apply the version of Portmanteau Theorem (weak convergence of measures)[2] to obtain:

$$\inf_{\pi\in\Pi(\mu,\nu)}K(\pi)=\lim_{n\to\infty}\int_{X\times Y}c(x,y)\,d\pi_n(x,y)\geq\int_{X\times Y}c(x,y)\,d\pi^*(x,y).$$

□

3 Kantorovich Duality

Now that we know the solution exists, how do we find a solution? Duality is the key. To construct our dual problem, we first rewrite the Kantorovich problem as an unconstrained optimization problem.

We will make use of the common trick in convex optimization that allows us to formulate the dual problem.

For C a convex set and f a convex function, we have that

$$\min_{x\in C}f(x)=\min_{x\in\mathbb{R}^n}f(x)+\begin{cases}0, & x\in C\\ +\infty, & \text{otherwise.}\end{cases}=\min_{x\in\mathbb{R}^n}f(x)+I_C(x).$$

Using a similar reasoning for the Kantorovich problem we can define the indicator function by

$$I_{\Pi(\mu,\nu)}(\pi)=\begin{cases}0, & \text{if } \pi\in\Pi(\mu,\nu)\\ +\infty, & \text{else,}\end{cases}$$

and let $\mathcal{M}_+$ denote the set of non-negative Borel measures on $X\times Y$. Then,

$$\inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)=\inf_{\pi\in\Pi(\mu,\nu)}\int c(x,y)\,d\pi=\inf_{\pi\in\mathcal{M}_+}\left(\int c(x,y)\,d\pi+I_{\Pi(\mu,\nu)}(\pi)\right)\quad\ldots\quad(*),$$

Now noting that,

$$I_{\Pi(\mu,\nu)}(\pi)=\sup_{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)}\left\{\int\varphi\,d\mu+\int\psi\,d\nu-\int(\varphi(x)+\psi(y))\,d\pi(x,y)\right\},$$

[2]For $\pi_n\to\pi$, weakly, the $\liminf_{n\to\infty}\int f d\pi_n\geq\int f d\pi\quad\forall f$, bounded from below and lower semicontinuous.

gives,

$$
\begin{aligned}
(*) &= \inf_{\pi\in\mathcal{M}_+}\sup_{(\varphi,\psi)}\left\{\int_{X\times Y}c(x,y)d\pi(x,y)+\int_X\varphi d\mu+\int_Y\psi d\nu-\int_{X\times Y}(\varphi(x)+\psi(y))d\pi(x,y)\right\}\\
&\underset{(1)}{=}\sup_{(\varphi,\psi)}\inf_{\pi\in\mathcal{M}_+}\left\{\int_{X\times Y}c(x,y)d\pi(x,y)+\int_X\varphi d\mu+\int_Y\psi d\nu-\int_{X\times Y}(\varphi(x)+\psi(y))d\pi(x,y)\right\}\\
&=\sup_{(\varphi,\psi)}\left\{\int_X\varphi d\mu+\int_Y\psi d\nu-\sup_{\pi\in\mathcal{M}_+}\int_{X\times Y}\{(\varphi(x)+\psi(y)-c(x,y))d\pi(x,y)\}\right\}\quad\ldots\quad(**)
\end{aligned}
$$

Note that for (1), we assumed a minimax principle to switch the infimum and supremum[3]. Now, suppose that $\varphi(x)+\psi(y)-c(x,y)=\epsilon>0$ for some (x_0,y_0). Then, we can define $\pi=\lambda\delta_{(x_0,y_)}$ and take $\lambda\to\infty$, which would imply that

$$\sup_{\pi\in\mathcal{M}_+}\int_{X\times Y}(\varphi(x)+\psi(y)-c(x,y))d\pi(x,y)\geq\lambda\epsilon\to\infty.$$

If $\varphi(x)+\psi(y)-c(x,y)\leq 0$ a.e. (i.e. $\mu-$a.e $x\in X$ and $\nu-$a.e $y\in Y$) the supremum is obtained at $\pi=0$. Then, the set of feasible solutions for the dual is given by:

$$D(\mu,\nu)=\{(\varphi,\psi)\in L^1(\mu)\times L^1(\nu):\varphi(x)+\psi(y)\leq c(x,y)\}.$$

Thus,

$$\inf_{\pi\in\mathcal{M}_+}\int_{X\times Y}(\varphi(x)+\psi(y)-c(x,y))d\pi(x,y)=\begin{cases}0, & (\varphi,\psi)\in D(\mu,\nu),\\ +\infty, & \text{otherwise.}\end{cases}$$

Therefore the Kantorovich optimal transport problem can be formulated as:

$$\inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)=\inf_{\pi\in\Pi(\mu,\nu)}\int_{X\times Y}c(x,y)d\pi=\sup_{(\varphi,\psi)\in D(\mu,\nu)}\left\{\int_X\varphi d\mu+\int_Y\psi d\nu\right\}=\sup_{(\varphi,\psi)\in D(\mu,\nu)}\mathbb{J}(\varphi,\psi),$$

Where $\mathbb{J}:L^1(\mu)\times L^1(\nu)\to\mathbb{R}$.

Theorem 3.1 (Kantorovich Duality). *Let $\mu\in\mathcal{P}(X)$, $\nu\in\mathcal{P}(Y)$ where X,Y are Polish spaces. Let $c:X\times Y\to[0,+\infty)$ be a lower semi-continuous cost function. Then (using the above notation),*

$$\min_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)=\sup_{(\varphi,\psi)\in D(\mu,\nu)}\mathbb{J}(\varphi,\psi).$$

Although strong duality formally relates to the values of objective functions, it is often observed that the solutions to the primal and dual problems are closely interconnected.

[3]To make this statement rigurous we can rely on the **Fenchel-Rockafellar Duality** Theorem, see [4].

3.1 Existence of solutions to the Dual Problem

The theorem stated in this section and its proof are based on [5] and [6].

Definition 3.1 (c-transform). *Let $\varphi\colon X \to \mathbb{R}\cup\{\infty\}$, the c-transform of φ is denoted by $\varphi^c\colon Y \to \mathbb{R}\cup\{\infty\}$ and defined by*

$$\varphi^c\colon Y \to \mathbb{R} \quad \varphi^c(y) = \inf_{x\in X}\{c(x,y) - \varphi(x)\}.$$

Theorem 3.2 (Existence of maximizers). *Let $\mu \in P(X)$, $\nu \in P(Y)$, where X and Y are Polish Spaces, and $c\colon X \times Y \to [0,\infty)$. Assume that there exists $c_X \in L^1(\mu)$, $c_Y \in L^1(\nu)$ such that $c(x,y) \le c_X(x) + c_Y(y)$ for μ-almost every $x \in X$ and ν-almost every $y \in Y$. In addition, assume that*

$$M := \int_X c_X(x)\,d\mu(x) + \int_Y c_Y(y)\,d\nu(y) < \infty. \tag{1}$$

Then there exists $(\varphi,\psi) \in D(\mu,\nu)$ such that

$$\sup_{D(\mu,\nu)} J = J(\varphi,\psi).$$

Furthermore we can choose $(\varphi,\psi) = (\eta^{cc},\eta^c)$ for some $\eta \in L^1(\mu)$.

To prove this theorem, we will make use of two lemmas. Essentially, the first lemma suggests that we only need to consider c-transform pairs instead of all $(\varphi,\psi) \in D(\mu,\nu)$. The second lemma aids in establishing upper bounds on maximizing sequences, which will be used to prove the existence of a maximizing function.

Lemma 3.1. *Let $\mu \in P(X)$, $\nu \in P(Y)$. For any $a \in \mathbb{R}$, and $(\tilde\varphi,\tilde\psi) \in D$ we have $(\varphi,\psi) = (\tilde\varphi - a, \tilde\psi + a)$ satisfies $J(\varphi,\psi) \ge J(\tilde\varphi,\tilde\psi)$ and $\varphi(x) + \psi(y) \le c(x,y)$ for μ-almost every $x \in X$ and ν-almost every $y \in Y$. Furthermore, if $J(\tilde\varphi,\tilde\psi) > -\infty$, $M < +\infty$, then there exists $C_X \in L^1(\mu)$ and $C_Y \in L^1(\nu)$ such that $(\varphi,\psi) \in D$.*

To be added—-

Proof. To be added. □

Lemma 3.2.

Proof. □

With the above two lemmas we are ready to prove the existence theorem of maximizers for the dual problem.

Proof. of Theorem (Existence of maximizers). To be added. □

From now on, we will focus our analysis on the quadratic cost, $c(x,y) = \|x-y\|_2^2$. Before that, we will give a brief review of convexity that will be useful in subsequent sections.

4 Review of Convexity

Definition 4.1 (Fenchel conjugate). *The convex conjugate of a proper function* $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is defined by*

$$\varphi^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y\rangle - \varphi(x)\}.$$

The subdifferential is a generalization of the differential that always exists for lower semi-continuous convex functions.

Definition 4.2 (Subdifferential). *Let* $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be a convex function. The subdifferential of* f *at a point* $x \in \mathbb{R}^n$*, denoted by* $\partial f(x)$*, is defined as the set*

$$\partial f(x) = \{v \in \mathbb{R}^n \mid \forall y \in \mathbb{R}^n, f(y) \geq f(x) + v \cdot (y - x)\}.$$

Proposition 4.1 (Subdifferential characterization). *Let* φ *be a proper, lower semi-continuous, convex function on* $\mathbb{R}^n$*. Then for all* $x, y \in \mathbb{R}^n$

$$x \cdot y = \varphi(x) + \varphi^*(y) \iff y \in \partial\varphi(x).$$

Proposition 4.2. *If* $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is convex then (1)* φ *is almost everywhere differentiable and (2) whenever* φ *is differentiable* $\partial\varphi(x) = \{\nabla\varphi(x)\}$*.*

Proposition 4.3 (Legendre duality for l.s.c convex functions). *Let* $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be proper. Then the following are equivalent:*

1. φ *is convex and lower semi-continuous;*
2. $\varphi = \psi^*$ *for some proper function* ψ*;*
3. $\varphi^{**} = \varphi$*.*

5 Case: Quadratic cost

We will restrict the following analysis to the case where $X, Y \in \mathbb{R}^n$ and $c(x, y) = \|x - y\|_2^2$. In turns out that the quadratic cost have close connection with Convex Analysis. First, we restate our Kantorovich problem in an equivalent and convenient manner.

Recall than in the dual problem we were constrained to the space

$$D(\mu, \nu) = \{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) : \varphi(x) + \psi(y) \leq c(x, y)\}.$$

Thus, $(\varphi, \psi) \in D(\mu, \nu)$ implies $\varphi(x) + \psi(y) \leq \frac{\|\mathbf{x}-\mathbf{y}\|^2}{2}$ $\mathrm{d}\mu$-a.e. and $\mathrm{d}\nu$-a.e. This leads to

$$\varphi(\mathbf{x}) + \psi(\mathbf{y}) \leq \frac{\|\mathbf{x}\|^2}{2} + \frac{\|\mathbf{y}\|^2}{2} - \langle \mathbf{x}, \mathbf{y}\rangle \iff \langle \mathbf{x}, \mathbf{y}\rangle \leq \left(\frac{\|\mathbf{x}\|^2}{2} - \varphi(\mathbf{x})\right) + \left(\frac{\|\mathbf{y}\|^2}{2} - \psi(\mathbf{y})\right).$$

Define, $\frac{\|\mathbf{x}\|^2}{2} - \varphi(\mathbf{x}) = \tilde{\varphi}$ and $\frac{\|\mathbf{y}\|^2}{2} - \psi(\mathbf{y}) = \tilde{\psi}$. Clearly $\tilde{\varphi} \in L^1(\mu)$, $\tilde{\psi} \in L^1(\nu)$ whenever μ, ν have finite second moments. Let

$$M_2 := \int_{\mathbb{R}^n} \frac{\|x\|^2}{2}\, d\mu(x) + \int_{\mathbb{R}^n} \frac{\|y\|^2}{2}\, d\nu(y) < +\infty.$$

Then we can rewrite our primal problem as

$$\inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int \frac{\|x-y\|^2}{2}\, d\pi(x,y) \right\} = M_2 - \sup_{\pi \in \Pi(\mu,\nu)} \left\{ \int \langle x, y \rangle\, d\pi(x,y) \right\},$$

and the dual problem as

$$\sup_{D} \left\{ \int \varphi\, d\mu + \int \psi\, d\nu \right\} = M_2 - \inf_{\tilde{D}} \left\{ \int \tilde{\varphi}\, d\mu + \int \tilde{\psi}\, d\nu \right\},$$

where $\tilde{D} := \left\{ (\tilde{\varphi}, \tilde{\psi}) \in L^1(\mu) \times L^1(\nu) : \langle x, y \rangle \leq \tilde{\varphi}(x) + \tilde{\psi}(y) \right\}$. Thus Kantorovich Duality becomes:

$$\sup_{\pi \in \Pi} \left\{ \int \langle x, y \rangle\, d\pi(x,y) \right\} = \inf_{\tilde{D}} \left\{ \int \tilde{\varphi}\, d\mu + \int \tilde{\psi}\, d\nu \right\}$$

The following Remark is the analogue of Lemma 3.1, which is the lemma that states that we only need to restrict the minimization problem to a certain type of functions in in $\tilde{D}$.

Remark 5.1.

$$\inf_{(\tilde{\varphi},\tilde{\psi}) \in \tilde{D}} \left\{ \int \tilde{\varphi}\, d\mu + \int \tilde{\psi}\, d\nu \right\} \geq \inf_{\varphi \in L^1(\mu)} \left\{ \int \varphi^{**}\, d\mu + \int \varphi^{*}\, d\nu \right\}$$

Proof. The result follows by noting that:

$$J(\tilde{\varphi}, \tilde{\psi}) \underset{(1)}{\geq} J(\tilde{\varphi}, \tilde{\varphi}^*) \underset{(2)}{\geq} J(\tilde{\varphi}^{**}, \tilde{\varphi}^*)$$

where

(1) $\tilde{\varphi}^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y \rangle - \tilde{\varphi}(x)\} \underset{(\tilde{D})}{\leq} \tilde{\psi}(y)$

(2) Note $\langle x, y \rangle \leq \tilde{\varphi}(x) + \tilde{\varphi}^*(y)$ so $(\tilde{\varphi}, \tilde{\varphi}^*) \in \tilde{D} \implies \tilde{\varphi}^{**}(x) = \sup_{y \in \mathbb{R}^n} \{\langle x, y \rangle - \tilde{\varphi}^*(y)\} \leq \tilde{\varphi}(x)$

Therefore,

$$\inf_{(\tilde{\varphi},\tilde{\psi}) \in \tilde{D}} \left\{ \int \tilde{\varphi}\, d\mu + \int \tilde{\psi}\, d\nu \right\} \geq \inf_{\varphi \in L^1(\mu)} \left\{ \int \varphi^{**}\, d\mu + \int \varphi^{*}\, d\nu \right\}$$

□

Therefore, minimizers of the dual problem take the form $(\varphi^{**}, \varphi^*)$, that is we can take infimum restricted to pairs $(\varphi^{**}, \varphi^*)$, which are convex lower semicontinuous functions.

Theorem 5.1 (Existence of an optimal pair of convex conjugate functions). *Let μ, ν be two probability measures on $\mathbb{R}^n$ with finite second moment. Then, there exists a pair (φ, φ^*) of l.s.c. proper conjugate convex functions on $\mathbb{R}^n$ such that:*

$$\inf_{\tilde{D}} \left\{ \int \tilde{\varphi}\, d\mu + \int \tilde{\psi}\, d\nu \right\} = \int \varphi\, d\mu + \int \varphi^*\, d\nu.$$

Proof. This is just a corollary of Theorem 3.2 with the necessary adaptations to the quadratic cost, see Theorem 1.14 in [3] for one equivalence. □

5.1 Optimality Criterion with Quadratic Cost (Knott-Smith)

When a transport plan is optimal? Knott-Smith Optimality criterion states that a transfer plan is optimal if and only if it is concentrated the sub-differential of a convex function.

Theorem 5.2 (Knott-Smith Optimality). *Let $\mu \in P(X)$, $\nu \in P(Y)$ with $X, Y \subseteq \mathbb{R}^n$ and assume that μ, ν both have finite second moments. A plan $\pi^* \in \Pi(\mu, \nu)$ is optimal if and only if there exists a convex lower semi-continuous function φ such that for π-almost all (x, y), $y \in \partial\varphi(x)$. Then, (φ, φ^*) is a minimizer of the problem $\inf_{\tilde{D}(\mu,\nu)} \mathbb{J}(\varphi, \psi)$, that is:*

$$\inf_{\tilde{D}(\mu,\nu)} \mathbb{J}(\varphi, \psi) = \int_X \varphi(x)\, d\mu(x) + \int_Y \varphi^*(y)\, d\nu(y)$$

where $\langle x, y \rangle \leq \varphi(x) + \varphi^*(y)$ for all x, y.

Proof. Let $\pi^* \in \Pi(\mu, \nu)$ and $\tilde{\varphi}$ be the proper lower semi-continuous function such that the pair $(\tilde{\varphi}, \tilde{\varphi}^*)$ minimizes J over $\tilde{D}$. The, by duality, we have

$$\int_X \tilde{\varphi}(x)\, d\mu(x) + \int_Y \tilde{\varphi}^*(y)\, d\nu(y) = \int_{X\times Y} x \cdot y\, d\pi^*(x, y).$$

Equivalently,

$$\int_{X\times Y} \tilde{\varphi}(x) + \tilde{\varphi}^*(y) - x \cdot y\, d\pi^*(x, y) = 0.$$

We must have $\tilde{\varphi}(x) + \tilde{\varphi}^*(y) = x \cdot y$ for π^*-almost every (x, y), and therefore $y \in \partial\tilde{\varphi}(x)$ for π^*-almost every (x, y).
Conversely, suppose $y \in \partial\tilde{\varphi}(x)$ for π^*-almost every (x, y) where $\tilde{\varphi}$ is a $L^1(\mu)$ proper, lower semi-continuous and convex function. Then,

$$\int_{X\times Y} \tilde{\varphi}(x) + \tilde{\varphi}^*(y) - x \cdot y\, d\pi^*(x, y) = 0.$$

We leave the integrability of $\tilde{\varphi}^*$ to the reader, then $(\tilde{\varphi}, \tilde{\varphi}^*) \in \tilde{D}$. Hence,

$$\min_{\tilde{\varphi}} J \leq J(\tilde{\varphi}, \tilde{\varphi}^*) = \int_{X\times Y} x \cdot y \, d\pi^*(x,y) \leq \max_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} x \cdot y \, d\pi(x,y).$$

By duality it follows that $(\tilde{\varphi}, \tilde{\varphi}^*)$ achieves the minimum of J and π^* achieves the maximum of $\int_{X\times Y} x \cdot y \, d\pi(x,y)$ in $\Pi(\mu,\nu)$. Hence π^* is an optimal plan in the Kantorovich sense.

□

5.2 Brenier's Theorem

The following theorem, established by Brenier, states that under mild regularity conditions, the solution to the Kantorovich problem with quadratic cost is achieved by a deterministic coupling, which is unique.

Theorem 5.3 (Brenier's Theorem). *Let $\mu \in P(X)$, $\nu \in P(Y)$ with $X, Y \subseteq \mathbb{R}^n$ and assume that μ, ν both have finite second moments and that μ does not give mass to small sets. Then there is a unique solution $\pi^* \in \Pi(\mu,\nu)$ to Kantorovich's optimal transport problem with cost $c(x,y) = \frac{1}{2}\|x-y\|^2$ which is given by*

$$\pi^* = (Id \times \nabla\varphi)_{\#}\mu$$

where $\nabla\varphi$ is the gradient of a convex function (defined μ-almost everywhere) that pushes μ forward to ν, i.e., $(\nabla\varphi)_{\#}\mu = \nu$.

Proof. Let π^* be a minimizer of Kantorovich's optimal transport problem.

By Knott-Smith, there exists a convex lower semi-continuous function φ such that for π^*-almost every (x,y), $y \in \partial\varphi(x)$. But convexity implies that φ is λ-almost everywhere differentiable ($\mu << \lambda$, then μ-a.e. diff.) Since μ doesn't assign mass to small sets, "All mass is concentrated in the interior of the domain." Therefore, μ-almost everywhere $\partial\varphi(x) = \{\nabla\varphi(x)\}$ and therefore μ-almost everywhere $\nabla\varphi(x) = y$ for π^*-almost every (x,y).

Note, if $y = \nabla\varphi(x)$ almost everywhere (a.e.) then $\pi = (\mathrm{Id} \times \nabla\varphi)_{\#}\mu$.

$$\begin{aligned} V(B) &= \pi^*(\mathbb{R}^n \times B) = \mu\left((\mathrm{Id} \times \nabla\varphi)^{-1}(\mathbb{R}^n \times B)\right) \\ &= \mu\left(\{x : (\mathrm{Id} \times \nabla\varphi)(x) \in \mathbb{R}^n \times B\}\right) \\ &= \mu\left(\{x : \nabla\varphi(x) \in B\}\right) \\ &= \mu\left(\nabla\varphi^{-1}(B)\right) = (\nabla\varphi)_{\#}\mu(B). \end{aligned}$$

It remains to show uniqueness of π^*.
Suppose there exists f such that $(\nabla f)_{\#}\mu = \nu$. We'll show $\nabla\varphi = \nabla f$, μ-a.e.. (i.e., the transport maps are the same)

By Knott-Smith: (f, f^*) is optimal for the dual:

$$\int f\,d\mu + \int f^*\,d\nu = \int \varphi\,d\mu + \int \varphi^*\,d\nu$$

Let π^* be associated to (φ, φ^*). Then,

$$\int (f + f^*)\,d\pi^* = \int (\varphi + \varphi^*)\,d\pi^* = \int \langle x, y\rangle\,d\pi^*,$$

Now, since $\pi^* = (\mathrm{Id} \times \nabla\varphi)_{\#}\mu$, we have:

$$\Rightarrow \int f(x)\,d\mu + \int f^*(\nabla\varphi(x))\,d\mu = \int \langle x, \nabla\varphi(x)\rangle\,d\mu \Rightarrow \int \left(f(x) + f^*(\nabla\varphi(x)) - \langle x, \nabla\varphi(x)\rangle\right)\,d\mu = 0$$

Recall $(f, f^*) \in \tilde{D}$,

$$\Rightarrow f(x) + f^*(\nabla\varphi(x)) = \langle x, \nabla\varphi(x), \rangle \quad \mu\text{-a.e.}$$
$$\therefore \nabla\varphi(x) \in \partial f(x), \quad \mu\text{-a.e. } x.$$

But f is μ-a.e. differentiable , thus,

$$\nabla\varphi(x) = \nabla f(x), \quad \mu\text{-a.e. } x$$

□

6 Wasserstein distance

Definition 6.1 (p-Wasserstein distance). *The* p-Wasserstein distance *between two probability measures μ and ν is defined via*

$$W_p^p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p\,d\pi(x, y),$$

The space of probability measures on $\mathbb{R}^d$ with finite p-th moment is denoted by

$$\mathcal{P}_p(\mathbb{R}^d) := \left\{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|^p\,d\mu(x) < \infty\right\}.$$

The space of probability measures with the p-Wasserstein distance can be viewed as a Riemannian Manifold.

Definition 6.2 (Riemannian Manifold). *A Riemannian Manifold $(M, \langle\cdot,\cdot\rangle)$ is a real smooth manifold M (differential manifold locally similar to a vector space) equipped with a positive-definite inner product $\langle\cdot,\cdot\rangle_p$ on the tangent space $T_p(M)$ at each point p.*

Proposition 6.1. *The space $(\mathcal{P}_p(\mathbb{R}^d), W_2)$ is a metric space. That is p-Wasserstein distance defines a metric over $\mathcal{P}_p(\mathbb{R}^d)$, that is for all $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ it holds:*

1. $W_p(\mu, \nu) \geq 0$
2. $W_p(\mu, \nu) = W_p(\nu, \mu)$
3. $W_p(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$
4. $W_p(\mu, \nu) \leq W_p(\mu, \xi) + W_p(\xi, \nu)$ *for all* $\xi \in \mathcal{P}_p(\mathbb{R}^d)$.

Remark 6.1. *Wasserstein distances induce a useful topology on r.v.'s. They metrize weak convergence on compact spaces.*

$$W_p(\mu_n, \mu) \to 0 \Leftrightarrow \mu_n \rightharpoonup \mu \quad \text{and} \quad \int \|\cdot\|^p \, d\mu_n \to \int \|\cdot\|^p \, d\mu.$$

As previously observed, W_p for $p = 2$ exhibits a remarkably special structure. Specifically for $p = 2$, there is a close connection with Convex Analysis.
The following theorem, sometimes referred to as the fundamental theorem of optimal transport, see [2], for the 2-Wasserstein distance is a direct consequence of the aforementioned results.

Theorem 6.1 (Fundamental theorem of optimal transport). *Let $\mu, \nu \in P_2(\mathbb{R}^d)$. Then, the following assertions hold.*

1. *(strong duality) The value of the dual optimal transport problem from μ to ν equals $\frac{1}{2}W_2^2(\mu, \nu)$.*
2. *(existence of optimal dual potentials) There exists an optimal pair (f^*, g^*) for the dual optimal transport problem.*
3. *(characterization of optimality) The optimal dual potentials are of the form*

$$f^* = \frac{\|\cdot\|^2}{2} - \varphi, \quad g^* = \frac{\|\cdot\|^2}{2} - \varphi^*,$$

   where $\varphi \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex, lower semicontinuous function and φ^* is its convex conjugate. If f^* denotes the optimal transport plan, then for f^*-almost every $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, it holds that $\varphi(x) + \varphi^*(y) = \langle x, y \rangle$, i.e., f^* is supported on the subdifferential of φ.
4. *(Brenier's theorem) Suppose in addition that μ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d$. Then, the optimal transport plan is unique, and moreover it is induced by an optimal transport map T. The mapping T is characterized as the (μ-almost surely) unique gradient of a proper convex lower semicontinuous function φ which pushes forward μ to ν: $T = \nabla\varphi$ and $(\nabla\varphi)_{\#}\mu = \nu$.*

References

[1] L. Ambrosio and N. Gigli. *A User's Guide to Optimal Transport*. In *Modeling and Optimisation of Flows on Networks*, Lecture Notes in Mathematics, Springer, 2013.

[2] S. Chewi. *Log-Concave Sampling*. Unfinished draft, April 18, 2024.

[3] S. Chewi, J. Niles-Weed, P. Rigollet. *Statistical Optimal Transport*. March 27, 2024.

[4] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

[5] M. Thorpe. *Introduction to Optimal Transport*. Centre for Mathematical Sciences, University of Cambridge, Lent 2018.

[6] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[7] C. Villani. *Optimal Transport: Old and New*. Springer-Verlag, 2009.

[8] K. Vodrahalli. *Geometry of Optimal Transport*. Simons OT Seminar, June 14, 2019.