University of British Columbia

Senior Honors Thesis

Some Results in the Theory of Optimal Transportation

Author:
Mohammad Bavarian

Supervisor:
Prof. Young-Heon Kim

April 25, 2011

Contents

1 Introduction

Assume we are given two mass distributions μ and ν on some space for example $\mathbb{R}^n$, one representing the distribution of a production of some good and the other the distribuion of its consumption, and we are asked to find the most efficient way to transfer the goods from the production locations to their desired locations of consumption. Of course, to solve this problem in any meaningful way we need information on the cost of tranportation between points of our space. In most applications, $c(x, y)$ the cost is taken to be a postive valued increasing function of distance

$$c(x, y) = f(|x - y|)$$

In the typical case that our original space is euclidean $|x - y|$ is just the euclidean distanc between $x, y \in \mathbb{R}^n$. For a Riemannian manifold (M, g) this is replaced by distance induced by the metric.

The formal format of the optimal tranport problem in euclidean space, in classical sense of *Monge*, then takes the following shape: Given two finite, possibly compactly supported, Borel measures μ and ν over $\mathbb{R}^n$ with equal total mass $\mu(\mathbb{R}^n) = \nu(\mathbb{R}^n) < \infty$ and Borel measurable cost $c(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+ \cup \{\infty\}$, we are interested in existence, uniqueness and regularity of the Borel mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ satisfying

$$\int_A d\nu = \int_{T^{-1}(A)} d\mu \qquad \forall A \in \mathcal{B}_{\mathbb{R}^n}$$

and minimizing the cost functional

$$I(T) := \int_{\mathbb{R}^n} c(x, T(x)) d\mu$$

over the set of all such T. The first condition here basically says that T pushes forwards, i.e. transports, the measure μ to ν and the second condition is the condition of optimality. Fixing the cost notice that this setting gives a distance between two such measures μ and ν. Monge problem is quite nonlinear in nature. To see this assume μ and ν are absolutely continuous with respect to Lebesgue measure, and hence are given by $d\mu = f\, dm$ and $d\nu = g\, dm$. Now let us try to solve the problem in the class of C^1 diffeomorphims. Then using the change of variable formula the push forward condition reads as

$$g(T(x))\ |\det \nabla T(x)|\ =\ \pm f(x) \tag{1}$$

which is a nonlinear equation different from our usual elliptic, parabalic, hyperbolic PDEs. It is hopeless to try tackling the optimal tranport problem too generally as the following negative results illustrate:

Example 1 *When there are point masses in our distribution the Monge question could break down. For example if* $\mu = \delta_0$ *is the Dirac delta measure at 0 and* $\nu = \frac{\delta_{-1}+\delta_1}{2}$*, then Monge equation has no solution because any push forward of the single point concenterated measure of* μ *is going to be single point concenterated, and hence there is no map that pushes* μ *forward to* ν*.*

Example 2 *In the space* $\mathbb{R}^2$ *take* μ *to be the restriction of* $\mathcal{H}^1$ *to* $\{0\}\times[0,1]$*. Let* ν *be restriction of* $(\frac{1}{2})\mathcal{H}^1$ *to* $\{-1,1\}\times[0,1]$*. The claim is that the optimal tranport problem between* μ *and* ν *with* $c(x,y) = \frac{1}{2}|x-y|^2$ *doesn't have any Monge solution. Indeed, as we shall see later there exist a unqiue generalized solution*[1] *to above problem. It is easy to characterize the optimal solution to this: In the optimal solution, the mass at each point is splitted into two equal parts transefered horizontally, one part to the left and the other to the right. This generalized solution is a multi-valued map and hence does not constitute a classical solution. There are in fact classical single valued tranfer plans that are almost optimal, i.e. their cost goes down as close as you desire to the optimal cost. To see this, you can split* $\{0\}\times[0,1]$ *into* $2N$ *intervals for a large* N *and tranfer even ones to right and the odd ones to left in the usual order.*

In light of these examples we shall assume from now on that our initial mass distribution μ is always absolutely continuous with respect to $\mathcal{L}^n$.[2] Also by scaling we can take μ and ν to be probability measures on $\mathbb{R}^n$. We'll denote the probability Borel measures on a space X by $\mathcal{P}(X)$. Also we'll use $T_\#\mu = \nu$ to say mapping T pushes measure μ forward to ν.

The most well understood version of optimal tranportation problem is the case of euclidean space with $c(x,y) = f(|x-y|)$ where f is either stricly convex or concave. It has been shown by Gangbo and Mccan among others [2] that in strictly convex case, assuming some further mild asymptotic conditions on the cost, the optimal tranport unqiuely exist. If some further assumptions are made [3] then existence and

[1] In the sense of Kantrovitch as introduced later.

[2] As it turns out this assumption can be relaxed. In most applications the assumption that μ would vanish on sets of Haussdorf dimension $n-1$ would be also sufficient.

[3] We must assume that ν vanishes on sets of Haussdorf dimnesion less than or equal to$n-1$ and, to have disjoint support away from μ. Of course without such assumptions counterexamples exist.f

uniquesness can be obtained for the stictly concave cost as well.

All approaches to the optimal tranportation known to us goes by first introducing a relaxed version of the problem, due to Kantorovitch. We shall present Kantorovitch prblem in generality, i.e. for two measure spaces X and Y with Borel measures.

Kantorovich Problem As example 2 shows in some cases splitting of mass must occur for the optimal solution to exist. Instead of looking for one-to-one mapping $T : X \to Y$ minimizing the cost functional, Kantrovitch allows the mass on one point to be distributed among all target points. More formally, Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ we want to find the probability measure γ on $X \times Y$ with marginals μ and ν i.e.

$$\mu(A) = \int_{A\times Y} d\gamma \qquad \forall A \in \mathcal{B}_X$$
$$\nu(B) = \int_{X\times B} d\gamma \qquad \forall B \in \mathcal{B}_Y$$

such that γ minimizes

$$I(\gamma) := \int_{X\times Y} c(x,y)\, d\gamma$$

The Kantrovitch problem is in a sense the linearized version of Monge problem. Unlike Monge problem, where the nonlinear condition of push forward makes it very hard to even establish existence of any solution, it is easy to establish existence results for Kantrovitch problem. This is because the space of probability measures with fixed marginals μ and ν has good compactness properties in weak topology. Furthermore, any mapping T of Monge $T : X \to Y$ with $T_{\#}\mu = \nu$ induces a measure $\gamma_T := (Id \times T)_{\#}\mu$ with marginal μ and ν satisfying

$$I(\gamma_T) = \int_{X\times Y} c(x,y) d\gamma_T = \int_X c(x, T(x)) d\mu = I(T)$$

which shows that the optimal solution to Kantrovitch problem is less costly than the optimal solution to the Monge problem.

Definition 1 *An optimal transference plan π to said to satisfy the* Monge property *if it is given by a Borel map $T : X \to Y$ as $\pi = (Id \times T)_{\#}\mu$.*

The usual approach to optimal tranportation problem is to first investigate existence, uniqueness and structure of the optimal solution to Kantorovich problem. In some cases then we are fotunate, as it turns out that the optimal solution of the Katrovitch γ satisfies the Monge property and hence gives us an exlicit optimal map T which solves Monge problem weakly. Regularity of γ and T then must be investigated to see whether this map T solves Monge equation in classical sense or not.

However, a reader not previously familiar with optimal tranportation theory, at this point might feel a little suspicious of our attempt to solve Monge problem by first solving the Kantrovitch problem. Actually it might be feared that the extra freedom which one has to split the mass into many (or infinitely many) pieces in Kantrovitch problem might enable one to conjure up a transfer plan which surpasses Monge optimal solution in cost efficiency. In that case, our attempt to solve Monge problem via first solving Kantrovitch problem is indeed hopeless. Indeed this could happen as we have already seen in example 2.

What went wrong in that example was the fact that the source mass distribution μ was concenterated in a set of dimension one less than the ambient space. In the same example if instead of having μ to be the restirction of 1-d measure $\mathcal{H}^1$ to $\{0\} \times [0,1]$, we spread the mass of μ in a neighborhood of this set by letting μ' be the restriction of $\frac{1}{2\epsilon}\mathcal{L}^2$ to $(-\epsilon,\epsilon) \times [0,1]$ the multivaluedness problem would have been fixed. Indeed for this μ' the solution to Kantrovitch problem would be a map that takes half of mass of μ' which is to the left of y-axis to $\{-1\} \times [0,1]$ and the other half to the right of y-axis to $\{1\} \times [0,1]$. Hence we see that when we spread the mass of source to avoid dimension problems the Kantrovitch solution became *almost μ-surely* single-valued. Consequently, Monge solution here will now coincide with Kantrovitch's solution.

So far we have investigated one difficulty that might prevent the Kantrovitch solution to become single-valued as to solve Monge problem, i.e. the problem of low dimension of support of source mass μ. The next example illustrates a second difficulty:

Example 3 *Let $c(x,y) = h(|x-y|)$ for some non-negative continuous function $h(y)$ that is strictly positive everywhere except at $y = 1$ where it takes value 0. Let μ be Lebesgue measure restricted to $(0,1)$ and ν be the restrcition of a half a Lebesgue measure to $(-1,0) \cup (2,3)$. It is clear that μ can be transfered to ν at 0 cost, which is of course optimal, by sending the mass at each point x to two points $x \pm 1$. Since $h(y) = 0$ implies $y = \pm 1$ it is easy to see that this map is the unique, up to μ-null sets, optimal map. So Kantrovitch's solution does not solve Monge's problem.*

In above example it is easy to check that Monge problem doesn't have any solution as the multi-valued plan above is the only possible zero cost plan. Similiar to construction in example2, by approximating the optimal plan we can produce near-optimal mappings. Indeed If we split $[0,1]$ to $2N$ equal intervals for a large N and we translate the odd intervals to the right and even ones to to the left, by continutiy of function h we get a map whose cost is very close to zero. So the infimum of the cost for Monge problem is also 0 but there is no optimal map for it, since we have seen the only zero cost map in above is Kantrovitch's multi-valued solution.

We need to understand what caused the failure in above example. First it must be noted that the assumption that $c(x,x) = h(0) \neq 0$ wasn't the cause of the problem as if you would have assumed $h(0)$ is zero in above would have made no difference. Indeed the same type of problem can occur anytime the cost function takes a mutiple local minimia. Here the two local minimas where at $r = \pm 1$. In such cases by an adaptation of above construction one can show the failure of existence of solution to Monge problem.

It is remarkable that the optimal solution to Kantrovitch problem can usually be approximated by single-valued mappings such that the cost becomes as close to the optimal cost as desired. We have already encountered this twice, once in the above discussion, and also in the example 2 of 1-dimensional source mass.

Indeed we have,

Proposition 1 *If μ and ν are probability measures over X and Y and μ has no atom, the space of viable solutions for Monge problem*[4] *is dense in the viable solutions for Kantrovitch problem. As a result, when the infimum cost is finite for optimal solution to Kantrovitch problem, one can always find single valued Monge transfer plans that come arbitrary close in cost to the optimal solution*

Sketch of Proof. The construction is exactly similiar to the examples. You start by covering the support of μ by countably many disjoint small cubes. We shall make cubes small enough such that $c(x,y) - c(x',y)$ be as small as some δ whenevever x, y belong to the same cube. For each cube P then you look at the image of P under optimal Kantrovitch transference plan γ

$$\nu_P(B) := \gamma(P \times B) \qquad \forall B \in \mathcal{B}_Y$$

Then to avoid to multivaluedness of original optimal γ it suffices that we rearrange the mapping from P to ν_P.[5] The rearragements in all cubes at most will cost us an extra $\mu[X] \cdot \delta$ which is small. Above arguement can be easily made precise.

[4] Of course here we only mean weak Monge problem, i.e. at the level of measure.

[5] The existence of such rearrangement is not so trivial to us.One way to obtain such a map is to

However, for this to be precise we need to be able to decompose the support of μ and ν to countably many campact pieces and in each compact piece $\{K_i\}_{i=1}^{\infty} \subset X$ and $\{H_i\}_{i=1}^{\infty} \subset Y$ exploit the uniform continuity of $c(x, y)$ in $K_i \times H_j$ to choose a uniform size for cubes and their target measure in that region. After this techincal work which uses the fact that μ and ν are regula, as all probablity measures on Polish speaces are, we will be done.□

We have seen a few examples where Kantrovitch's optimal solution hasn't been the solution to Monge problem. The remedy to the first problem we encountered was the new assumption that μ is absolutely continuous with respect to $\mathcal{L}^n$. What saves us from difficulties of second problem is to have strcitly increasing costs with some convexity-type[6] geometry. For example such assumptions already gets rid of the sort of diffculties we had in the example 3.

The major theme of this work is to study optimal tranportion for costs other than the well-established strictly convex and strictly concave cases. Through our study of such costs and our attempt to pinpoint the differences that the non-invertibility of ∇c makes, we will naturally encounter a new metric on the space of compactly supported probability measures $\mathcal{P}(\mathbb{R}^n)$. To our surprise this metric whose consideration was motivated by study of costs behaving differently in short-long ranges, relates back to the traditional theory of optimal transportation as it turns out to be the limit of traditional *p-Wasserstein distances*. We will study these concepts and prove some chracterization theorems. We will also state and motivate a few new conjectures and problems for further study.

2 Existence and Geometry of Optimal Transportations

Kantrovitch formulation of optimal tranport problem has the advantage that it allows one to obtain very general existence results. The ultimate reason behind such general existence theorems is some compactness result in the space of probability measures which combine with lower semiconituity of the map

take it to be the solution to optimal transportation of $\mu|_P$ to ν_P with the quadratic cost, which is single-valued by theory outlined in the next section. To work the details, regularity of ν and hence nu_P is crucial.

[6]convexity, concavity or saddle typre geometry of cost. Strictness of these type of geoemtries are often required in order to achieve unqiueness of solution

$$I : \pi \rightarrow \int_{X\times Y} c(x,y)\, d\pi(x,y)$$

to give us the existence of infimum, i.e. optimal Kantorovitch solution. After establishing this existence following the exposition of Villiani books[1, 4], We'll study the properties of the Kantovitch optimal solution. As it turns out, the support of such infimal π has a very nice geometric property called *c-cyclical monotonicity* which enables us to prove, following Gangbo and Mccan and Villani[1, 2], the single valuedness of the optimal transference plan π for stricly convex costs. This amounts to solving the Monge problem in a weak sense.

In this section it is not our intention to give full proofs and complte details to every statement. Our intention is to merely summarize and review the major results in the theory that are illuminating for the discussions in the subsequent sections. For more details reader should consult with excellent sources like [1, 4].

2.1 Existence of Optimal Plans

Let X and Y be polish spaces, i.e. complete seprable metric spaces. We want to establish the following theorem. Let μ and ν be probability measures on X and Y respectively. Denote by $\Pi(\mu,\nu)$ all probability measures on $X\times Y$ with marginals μ and ν. We have the following general theorem:

Theorem 1 . *Let $c(x,y)$ be a lower semi-contintuous non-negative Borel function over $X\times Y$. The there exist $\pi \in \Pi(\mu,\nu)$ minimizing the cost functional $I(\gamma) = \int_{X\times Y} c(x,y)\, d\gamma$ over all $\Pi(\mu,\nu)$.*

To prove this we first must show the compactness of $\Pi(\mu,\nu)$ in weak topology. The key fact is that this set is precompact. This is a consequence of *Prokhorov's theorem.* This theoreom states that for a Polish space X, a set $\mathcal{A} \subset \mathcal{P}(X)$ is precompact if and only if it is tight, i.e. for each $\epsilon > 0$ there exist a compact set K in X such that $\gamma[X\backslash K] < \epsilon$ for every γ in $\mathcal{A}$. See Villaini's book [4] and the refrences therein for the proof of this important result.

Now note that $X\times Y$ is indeed a Polish space because both X and Y are. Assume we are given an $\epsilon > 0$. Since every Borel probabality measure on a polish space is regular, So there is K and K' such that $\mu[X\backslash K] < \frac{\epsilon}{2}$ and $\nu[Y\backslash K'] < \frac{\epsilon}{2}$. Then we have

$$\gamma[X \times Y \backslash K \times K'] \leq \gamma[X \times Y \backslash K \times Y] + \gamma[X \times Y \backslash X \times K'] = \nu[Y \backslash K'] + \mu[X \backslash K] < \epsilon$$

Since X and Y are metric spaces $K \times K'$ is compact in $X \times Y$ (sequentially and hence generally by metric space property) and so tightness condition holds and by Prokhorov's theorem, $\Pi(\mu, \nu)$ is compact. Moreover, the conditions that the marginals are μ and ν is preserved by passing to a limit with weak convergence so $\Pi(\mu, \nu)$ is closed and precompact and so it is compact.

Proof of Thm 1 If $I(\gamma) = +\infty$ for all γ in $\Pi(\mu, \nu)$ then there is nothing to prove. So assume $I(\gamma) < \infty$ for some γ. Take a minimizing sequence π_k in $\Pi(\mu, \nu)$

$$\lim_{k \to \infty} I(\pi_k) = \inf_{\pi \in \Pi(\mu,\nu)} I(\pi)$$

By compactness and passing to a subsequence we can assume $\pi_k \to \pi$ for some π in $\Pi(\mu, \nu)$ as $k \to \infty$. Our goal is to show $I(\pi) = \lim_{k \to \infty} I(\pi_k) = \inf_{\pi \in \Pi(\mu,\nu)} I(\pi)$. Since c is a nonenegative lower semi-continuous function, we can easily find a sequence $\{c_m\}_{m=1}^{\infty}$ of continuous non-decreasing sequence of functions, uniformly bounded from below, that converge pointwise to c. Then by monotone convergence theorem and the fact that π_k go to π weakly we have,

$$\int c \, d\pi = \lim_{m \to \infty} \int c_m \, d\pi = \lim_{m \to \infty} \lim_{k \to \infty} \int c_m \, d\pi_k \leq \liminf_{k \to \infty} \int c \, d\pi_k$$

Indeed, the last expression is equal to $\inf_{\gamma \in \Pi(\mu,\nu)} I(\gamma)$ and hence π is indeed the minimizer $\square$

In above we relied in following lemma. Although it is not too difficult, we'll sketch its proof here.

Lemma 1 *Let f be a lower-semicontinuous nonnegative real-valued function over Polish space X. Then f can be approximated by a non-decreasing sequence of continuous functions converging up to f pointwise.*

Sketch of Proof. Let $E \subset X \times \mathbb{R}$ be the epigraph of f. Since E is closed E^{ϵ} is also a closed set.

$$E^{\epsilon} := \{p \in X \times \mathbb{R} \mid d(p, E) \leq \epsilon\}$$

Here $d(.,.)$ is the distance function induced from metric of X and $\mathbb{R}$ on the product space. Since E is closed, intersection of all E^{ϵ} for $\epsilon > 0$ is actually E itself.

Next observe the following fact : E^ϵ is epigraph of a function from X to $\mathbb{R}$ bounded from below by $-\epsilon$. The reason is if $P = (p, \lambda) \in X \times \mathbb{R}$ is in E^ϵ then $P' = (p, \lambda')$ is also in E^ϵ whenever $\lambda' > \lambda$. Indeed let $Q = (q, \alpha) \in E$ be distance at most ϵ to P. Then translate Q upward to get $Q' = (q, \alpha + (\lambda' - \lambda))$ which is alose in E and is distace at most ϵ to P'.

Hence, associated to each ϵ there is a lower semi-continuous function bounded below by $-\epsilon$ that has an epigraph E^ϵ. Now as ϵ goes to zero, $E^\epsilon \downarrow E$ which means the functions are non-decreasingly converge pointwise to f.

The final point is that the functions constructed are not only l.s.c. but also continuous. To see this notice that a l.s.c. function has a point of discontinuity if and only the closure of the interior of its epigraph is not whole of its epigraph. Now notice that every point that is distance less than ϵ to E is in interior of E^ϵ, and hence closure of E^ϵ is E^ϵ itself and so above functions are continuous. Now picking a sequence ϵ_n that go to zero and taking the associted function with epigraphs E^{ϵ_n} gives us our desired sequence. $\square$

2.2 Geometry of Optimal Transportation

Now we are going to investigate the propoerties of $\pi \in \Pi(\mu, \nu)$ the solution of Kantorovitch problem that we constructed in the previous part of this section. Following Gangbo and Mccan[2] we are going to work in an euclidean setting with a cost $c(x, y) = h(x-y)$, h strictly convex. The most important example of such costs for us are $d_p(x, y) = |x - y|^p$ for $p > 1$. Assuming some extra mild conditions on our stricly convex cost c which are satisfied by d_p's, and further assuming that μ is absolutely continuous with respect to Lebesgue measure, Gangbo and Mccan were able to show that the π constructed above actualy is of the form $\pi = (Id \times T)_{\#}\mu$ and hence solves the Monge problem. Furthermore, they proved uniqueness of such optimal mapping.

Here we shall not give proofs of the statements. We merely state the theorems relevent to our study and the idea behind some of the proofs. Interested reader should look at the sources like [1, 2] for the detailed account of the proofs.

First, Let us have X, Y Polish spaces and μ and ν probability measures on them as in previous part. We will later restrict to euclidean space. Assume a lower-semicontinuous cost function c.

Definition 2 *$\Gamma \subset X \times Y$ is called c-cyclically monotone if for any* $\{(x_i, y_i)\}_{i=1}^n \subset \Gamma$

we have

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

For all σ permution of $\{1, 2, \dots, n\}$

The relevancy of this definition to optimal map becomes more clear by the next theorem.

Theorem 2 *Assume the cost function c is continuous. The support of optimal measure $\pi \in \mathcal{P}(X \times Y)$ for Kantorovitch problem over $\Pi(\mu, \nu)$ has to be c-cycliccaly monotone subset of $X \times Y$.*

The idea of the proof is clear. If the cyclical monotonicity fails it means for some $\{(x_i, y_i)\}_{i=1}^{n} \subset Supp(\pi)$ and for some permutation σ we have

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

Now note that by continuity of c, above relation will still hold if you instead (x_i, y_i)'s and $(x_i, y_{\sigma(i)})$ you use points very close to them. So taking appropriately small balls B_i's around (x_i, y_i)'s and $\tilde{B}_i$ around $(x_i, y_{\sigma(i)})$, you can modify π to make it more cost efficient as follows: Since (x_i, y_i)'s were in the support, they all have non-zero π measure. Now if you carefully modify π by reducing it by a small amount on B_i's and adding an small amount to π on $\tilde{B}_i$ such that the modified π has the same marginals you will get a contradiction because you have just obtained a more efficient coupling of (μ, ν).

Corollary 1 *Take Γ to be the union of the support of all optimal π's solving Kantorovitch's problem in $\Pi(\mu, \nu)$. Then Γ is c-cyclically monotone set.*

We must show that for any finite set of points $\{(x_i, y_i)\}_{i=1}^{n} \subset \Gamma$ satisfy the cyclical monotonicity condition. Note that each (x_i, y_i) belongs to the support of some optimal γ_i. But then if you consider $\gamma = \frac{\gamma_1+\gamma_2+\ldots+\gamma_n}{n}$ it will have all (x_i, y_i)'s in its supports and it would be optimal in cost, and hence its support that includes our points is c-cycllically monotone. $\square$

Now we need to introduce a few new notions that generalize the familiar concavity and superdifferentiability of a function.

Definition 3 *A function $\psi : X \to \mathbb{R} \cup \{-\infty\}$ is called c-concave if there exist some function function $u : Y \to \mathbb{R} \cup \{-\infty\}$, $u \neq -\infty$ such that*

$$\psi(x) = \inf_{y \in Y} [c(x, y) - u(y)]$$

The domain of ψ, denoted by $\mathrm{dom}(\psi)$, is the set of point $x \in X$ where $\psi(x) \neq -\infty$.

Definition 4 *The c-superdifferential $\partial^c \psi$ of a c-concave function is the set of all pairs $(x, y) \in X \times Y$ such that*

$$\forall z \in X \qquad \psi(z) \leq \psi(x) + [c(z, y) - c(x, y)]$$

The reader is probably familiar with the notion of *Legendre transform* of a function. For a (convex) function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ not identically $+\infty$, Legendre transform tries to build a function $h^*(y)$ such that ∇h^* is the inverse of ∇h.[7]

$$h^*(y) := \sup_{x \in \mathbb{R}^n} [x.y - h(x)]$$

Now imitating the construction of Legendre transform, we'll introduce the notion of *c-transform* which tries to invert the c-subdifferential set. More precisely, for any function ψ on X with values in $\mathbb{R} \cup \{-\infty\}$ and not identically $-\infty$ define $\psi^c(y)$ on Y by,

$$\psi^c(y) \ := \ \inf_{x \in X} [c(x, y) - \psi(x)]$$

The importance of these notions are due to the following theorem.

Theorem 3 *Any $\Gamma \subset X \times Y$ c-cyclically monotone set lies in the $\partial^c \psi$ of some c-concave function.*

Now we shall assume that $X = Y = \mathbb{R}^n$ and $c(x, y) = h(x - y)$ with h strictly convex. we already saw that the union of supports of all optimal measures in $\Pi(\mu, \nu)$ is a c-cycllically monotone set. Using above theorem we arrive at a function ψ such that its c-subdifferential encompasses the whole of union of all optimal supports. Now assuming some other mild conditions, one can show that a c-concave function ψ must be Lipschitz continuous and hence almost everywhere $d\mu$ differentiable (Recall $d\mu$

[7] More precisely this invertibility property only holds when both function are differentiable. Also ∇h might not be invertible in which case ∇h^* is the best substitute for $(\nabla h)^{-1}$.

was assumed to be abs. continuous with respect to Lebesgue). As a consequence we must have

$$\partial^c\psi(x) = x - (\nabla h)^{-1}(\nabla\psi(x)) \qquad d\mu - \text{almost everywhere}$$

This is becuase $(x, y) \in \partial^c\psi$ means that $f(z) = \psi(x) + [c(z, y) - c(x, y)]$ becomes tangent to $\psi(z)$ at x which means that their derivatives must match at $z = x$. So we must have $\nabla\psi(x) = \nabla h(x - y)$. Applying $(\nabla h)^{-1}$ to this identity we get our desired above result which also shows that $\partial^c\psi(x)$ is single valued whenever ψ is differentiable at x. Also by Legendre duality we can replace $(\nabla h)^{-1}$ by ∇h^* in above.

Note that here we glossed over key important technical details. We didn't address the issues that why h^* is differentiable and why ψ is lipschitz. Gangbo and Mccan develop a regularity theory that proves these important points. For this they need some extra asymptotic conditions on $c(x, y) = h(x - y)$. We call such a cost an *optimally suitable cost.*

Definition 5 *A cost is called optimally suitable if $c(x, y) = h(x - y)$ with h strictly convex and superlinear $\lim_{|x|\to+\infty} \frac{h(|x|)}{|x|} = +\infty$ such that it satisfies the following condition: Given $r > 0$ and $\theta \in (0, \pi)$, whenever $p \in \mathbb{R}^n$ is far enough from the origin, there should exist a direction $z \in \mathbb{R}^n$ such that on the truncated cone $K(p, z, \theta)$ with angle $\frac{\theta}{2}$, and vertex p and direction z, defined by*

$$K(p, z, \theta) \equiv \left\{ x \in \mathbb{R}^n \quad | \quad |x - p||z|cos(\frac{\theta}{2}) \le \langle z, x - p\rangle \le r|z| \right\}$$

c assumes its maximum at p.

Now we can finally state the main theorem.

Theorem 4 *Let c be an optimally suitable cost and μ and ν Borel probablity measure on $\mathbb{R}^n$. If μ is absolutely continuous with respect to Lebesgue measure then there exist a c-concave potential ψ such that the map $s(x) = x - \nabla h^*(\nabla\psi)$ pushes μ forward to ν and mimizes the cost as a Kantorovitch problem minimizer over $\Pi(\mu, \nu)$. Furthermore, $(Id \times s)_\#\mu$ is the unique optimizer of Kantorovitch problem whenever the optimal cost is finite.*

This is our last theorem for this part and we shall not delve more deeply into the geometry and regularity of optimal maps for strictly convex costs. The above theorem for us would set for us the highest the standard for a solution to optimal tranportation problem. So in the next section when we move on the theory of costs behaving differently in short-long ranges and more generally theory of non-convex non-concave costs , the goal of our study would be to see how much of the above theorem proved in the class of strictly convex costs can be generalized to those settings. We also note that a similiar theorem can be proved for strictly concave costs if you assume μ and ν have disjoint supports.

2.3 More Major Results in Optimal Transportation

Here we shall talk about some more foundational concepts like *Kantorovitch duality*, *Wasserstein distances* and *displacement interpolation* which we must know to state and prove some of our major theorems and conjectures in the next section.

2.3.1 Kantorovitch Duality

Kantorovitch's problem is a linear problem with constraints. Linear programming tells us that linear problems with constraintes can be dualized. Such duality in our setting then would give rise to the following theorem.

Theorem 5 *Let X and Y be Polish spaces with μ and ν the usual source and target measures. And let the cost $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be lower-semicontinuous. For any $(\varphi, \psi) \in L^1(d\mu) \times L^1(d\nu)$ define $J(\varphi, \psi) = \int_X \varphi d\mu + \int_Y \psi d\nu$. Let Φ_c stand for all such pairs (φ, ψ) that satisfy $\varphi(x) + \psi(y) \leq c(x, y)$ for $d\mu$-almost all $x \in X$ and $d\nu$-almost all $y \in Y$. Then*

$$\inf_{\pi \in \Pi(\mu,\nu)} I(\pi) = \sup_{\Phi_c} J(\varphi, \psi)$$

The next theorem is where the most interesting information for us lies.

Theorem 6 *Given the assumptions of the previous theorem and the extra assumption that there exist nonegative functions $a(x) \in L^1(d\mu)$ and $b(y) \in L^1(d\nu)$ such that*

$$\forall (x, y) \in X \times Y \qquad c(x, y) \leq a(x) + b(x)$$

Then dual Monge problem $\sup_{\Phi_c} J(\varphi, \psi)$ has a maximizer in the form of a conjugate pair of c-concave functions. Furthermore for any (φ, ψ) pair maximizing the

Kantorovitch dual problem, any $\pi \in \Pi(\mu, \nu)$ minimizing the Kantorovitch original problem must be concenterated on the set of points

$$\{(x, y) \in X \times Y \mid \varphi(x) + \psi(y) = c(x, y)\}$$

This theorem is extremely interesting to us. If c is a complicated function understanding the shape and structure of conjugate pair of c-concave functions or even finding a simple criteria to characterize a single c-concave is extremely difficult. However, if the cost function is simple as our main cost function would be in the next section one might be tempted to chracterize exactly the structure c-concave conjugate pairs for such costs. The proof of above theorems can be found among other places in Villiani's texts [1, 4].

2.3.2 Introduction to Wasserstein Distances

Fix a Polish space X. Wasserstein distances are distances derived from optimal transportation on the space of probability measures $\mathcal{P}(X)$. Fix Borel probability measures μ and ν on X and let π_p denote an optimal transference plan for the cost $c(x, x') = d(x, x')^p$. Then we define W_p the Monge-Kantorovitch distance of order p,

$$W_p = \left(\int_{X \times X} |d(x, x')|^p \, d\pi_p \right)^{\frac{1}{p}}$$

Theorem 7 *For all $p \in [1, \infty)$, W_p defines a metric on the space $\mathcal{P}_p(X)$ which is the space of probability measures with finite moment of order p. Note that we way μ has finite p-th moment if these exist some $x_0 \in X$ such that*

$$\int_X d(x_0, x)^p \, d\mu < \infty$$

The fact that $W_p(\mu, \nu) = 0$ if and only if $\mu = \nu$ is clear. The moment condition guarantees that

$$\begin{aligned} W_p(\mu, \nu)^p = \int_{X \times X} d(x, x')^p \, d\pi_p &\leq 2^{p-1} \int_{X \times X} [d(x, x_0)^p + d(x', x_0)^p] \, d\pi_p \\ &\leq 2^{p-1} \int_{X \times X} [d(x, x_0)^p + d(x', x_0)^p] \, d\mu \, d\nu \\ &= 2^{p-1} \int_X d(x, x_0)^p \, d\mu + \int_X d(x', x_0)^p \, d\nu < +\infty \end{aligned}$$

So it only remains to check the triangle inequlity. The idea of the proof of triangle inequlity is clear but there is a little techinicality. We want to show

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

For transfering μ_1 to μ_2 we have an optimal transference plan γ_{12} that gives us a recpie of how to distribute the mass of the first measure to get the second measure. If we comibine this with the recpie γ_{23} that gives us the optimal way to distribute mass of the second measure the get the third, we arrive at a combined recpie γ_{13} that take the first measure to the third. For example, if $\gamma = (Id \times T_{12})_{\#}\mu_1$ and $\gamma' = (Id \times T_{23})_{\#}\mu_2$ then we can check that $\gamma_{13} = (Id \times T_{23} \circ T_{12})_{\#}\mu_1$. Then we expect to be able to establish using triangle and Holder inequality and the fact that $W_p(\mu_1, \mu_3)$ is optimal,

$$W_p(\mu_1, \mu_3) \leq \left(\int_{X \times X} d(x, x')^p \, d\gamma_{13} \right)^{\frac{1}{p}} \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3)$$

Now let's try to make this approach rigorous. For that we need the following important *Gluing Lemma.*

Lemma 2 ***(Gluing Lemma)*** *Let μ_1,μ_2 and μ_3 be three probability measures over Polish spaces X_1, X_2 and X_3. Let $\pi_{12} \in \Pi(\mu_1, \mu_2)$ and $\pi_{23} \in \Pi(\mu_2, \mu_3)$ be two transference plans. Then there exist a probability measure $\pi \in \mathcal{P}(X_1 \times X_2 \times X_3)$ with marginals π_{12} on $X_1 \times X_2$ and π_{23} on $X_2 \times X_3$.*

Now we will use this gluing lemma to finish the proof of triangle inequlity. Let $X_1 = X_2 = X_3 = X$. We already have π_{12} the optimal plan between μ_1 and μ_2, and π_{23} the optimal map between μ_2 and μ_3. Using gluing lemma we get $\pi \in \mathcal{P}(X_1 \times X_2 \times X_3)$. Let π_{13} projection of π on first and third factor, $X_1 \times X_3$. Then clearly $\pi_{13} \in \Pi(\mu_1, \mu_3)$. Using triangle and Minkowski inequalities we can write,

$$\begin{aligned}
W_p(\mu_1, \mu_3) &\leq \left(\int_{X_1 \times X_3} d(x_1, x_3)^p d\pi_{13}(x_1, x_3)\right)^{\frac{1}{p}} \\
&= \left(\int_{X_1 \times X_2 \times X_3} d(x_1, x_3)^p d\pi(x_1, x_2, x_3)\right)^{\frac{1}{p}} \\
&\leq \left(\int_{X_1 \times X_2 \times X_3} [d(x_1, x_2) + d(x_2, x_3)]^p d\pi(x_1, x_2, x_3)\right)^{\frac{1}{p}} \\
&\leq \left(\int_{X_1 \times X_2 \times X_3} d(x_1, x_2)^p d\pi(x_1, x_2, x_3)\right)^{\frac{1}{p}} + \left(\int_{X_1 \times X_2 \times X_3} d(x_2, x_3)^p d\pi(x_1, x_2, x_3)\right)^{\frac{1}{p}} \\
&= \left(\int_{X_1 \times X_2} d(x_1, x_2)^p d\pi(x_1, x_2)\right)^{\frac{1}{p}} + \left(\int_{X_2 \times X_3} d(x_2, x_3)^p d\pi(x_2, x_3)\right)^{\frac{1}{p}} \\
&= W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3)
\end{aligned}$$

We do not present a proof of Gluing Lemma here. It probably looks quite intuitive at the first sight but to do it generally, one needs to use the machinary of *disintegration of measure*. For a proof look at Villiani's book [1] and the references therein.

Finally we will say a few words regarding the topology induced on the space $\mathcal{P}_p(X)$ via the metric W_p. Take a sequence $(\mu_k)_{k \in \mathbf{N}}$ from $\mathcal{P}_p(X)$. There exist a $\mu \in \mathcal{P}(X)$ such that $W_p(\mu_k, \mu) \to 0$ if and only if $\mu_k \to \mu$ in the weak sense and also satisfy a *Tightness condition*: The tail of the p-th moment integral of μ_k's are uniformly small if the domain of tail is far away,

$$\lim_{R \to \infty} \limsup_{k \to \infty} \int_{d(x_0, x) \geq R} d(x_0, x)^p d\mu_k(x) = 0$$

Remark. In the next section we work in a situation when all W_p for $p > 1$ make sense and are finite, so we work in the space $\mathcal{P}_c(X)$ the space of compactly supported probability measures on X. Despite that all W_p exist over this space, the topologies aren't the same because the moment condition for $p > q$ is different and harder to satisfy for p. In fact, it is clear that the topology for $p > q$ is coarser becuase by Holder inequality

$$p_1 \geq p_2 \geq 1 \quad \Rightarrow \quad W_{p_1} \geq W_{p_2}$$

On the other hand if X is bounded with respect to d, which in our case implies we are restricting all our probability measures to K bounded subset of $\mathbb{R}^n$, then all the topologies of W_p is indeed the same because by interpolation we have,

$$p_1 \geq p_2 \geq 1 \quad \Rightarrow \quad W_{p_1} \geq W_{p_2}^{\frac{p_2}{p_1}} \mathrm{diam}(X)^{1-\frac{p_2}{p_1}}$$

2.3.3 Displacement Interpolation

In the section on geoemtery of optimal transportation we stated the main theorem of Gangbo and Mccan 4. We saw that given μ and ν absolutely continuous probability measures on $\mathbb{R}^n$ and $c(x,y) = h(x-y)$ optimally suitable cost, there exist a unique optimal map

$$T(x) = x - \nabla h^*(\nabla\varphi(x))$$

pushing μ to ν where $\varphi(x)$ is a c-concave function. Moreover, we noted that any map $T(x)$ of the above form when $\varphi(x)$ is c-concave is an optimal map between μ and $T_{\#}\mu$. We also mentioned that one can easily check that for $c_p(x,y) = |x-y|^p$, $p > 1$, is a optimally suitable cost. Now the interesting fact is that for $c_p(x,y) = |x-y|^p$ the property of being c-concave remains true under interpolating, i.e. if $\varphi(x)$ is c_p-concave, then $t\varphi(x)$ is also c_p-concave for any $t \in [0,1]$. This gives rise to the following fact :

Consider the Monge-Kantorovitch problem for $c_p = |x-y|^p$ for the euclidean space as descrived above. Then,

$$T_t(x) = x - \nabla h^*(\nabla t\varphi(x))$$

For each $t \in [0,1]$ the map T_t is an optimal map between μ and $T_{t\#}\mu$.

In above $T_1 = T$ so T_1 pushes μ to ν. On the other hand $T_0 = id$, so $T_{0\#}\mu$ is μ itself. The intermediates measure $T_{t\#}\mu = \mu_t$ interpolate between these two and hence the name *displacement interpolation* is appropriate.

Let q be the conjugate exponent for $p > 1$. Then $h_p^*(y) = C_p|y|^q$ for some constant C_p. So we see

$$T_t(x) = x - C_p\, t^q\, |\nabla\varphi(x)|^{q-1}\, \nabla\varphi(x)$$

So chainging the varible to $s = t^q$ we see for every $s \in [0,1]$ the map

$$T_s(x) = x - s\, C_p\; |\nabla\varphi(x)|^{q-1}\, \nabla\varphi(x)$$

is an optimal map.

Remark 1. When one varies the paremeter $s \in [0,1]$ the cost associated with T_s varies lineraly,

$$I(T_s) = s \cdot \ I(T)$$

Remark 2. The Monge property of $c_p(x, y)$ wasn't essential for defining the notion of displacement interpolation. Let $X = Y = \mathbb{R}^n$ be our setting of domains. We try here to generalize the *Displacement Interpolation property* in the case where the optimal map $\pi \in \Pi(\mu, \nu)$ does not apriori satisfy Monge property, i.e. the optimal transference plan might be multi-valued.

For any joint probability measure $\pi \in \mathcal{P}(X \times Y)$. By disintergration of measure we can write

$$\pi = \int_X (\delta_x \otimes \pi_x)\, d\mu(x)$$

The meaning of $\pi_x \in \mathcal{P}(Y)$ in above is the following: The map $x \to \pi_x$ associates to each x the probability measure π_x on Y which tells how the mass of μ at the point x splits among the points of Y. The integral $\int_X (\delta_x \otimes \pi_x) d\mu(x)$ basically says that π is just the average of all π_x weighted by the mass $d\mu(x)$. More precise meaning of above integral is that for any test function $\phi \in C_b(X \times Y)$

$$\int_{X \times Y} \phi(x, y)\, d\pi(x, y) = \int_X \left[\int_Y \phi(x, y) d\pi_x(y) \right] \, d\mu(x).$$

Now the s-contractiong for the case above was given by $T_s(x) = x - s(\nabla\varphi(x))^{q-1}\nabla\varphi(x)$ and under this s-contraction, the family T_s remained optimal. We can generalize this notion of s-contraction as follows to get the interpolating family π^s,

$$\pi^s \ := \int_X (\delta_x \otimes \pi_x^s)\, d\mu(x)$$

Where $\pi_x^s \in \mathcal{P}(Y)$ for $0 < s \leq 1$ the x-centered contraction of π_x is defined by the relation,

$$\pi_x[A] = \pi_x^s[s \cdot A + (1 - s) \cdot x] \qquad \forall A \in \mathcal{B}_Y$$

and $\pi_x^0 = \delta(x)$ for $s = 0$.

So we say our Monge-Kantorovitch problem satisfies the *Displacement interpolation* property if the s-concenterated familty π_s derived from optimal plan $\pi \in \Pi(\mu, \nu)$ remains optimal for the Kantorovitch problems for $\pi^s \in \pi(\mu, \mu_s)$

$$\mu_s[A] := \pi^s[X \times A] \qquad \forall A \in \mathcal{B}_Y$$

These two remarks highlight what we'll try to generalize when in the next section where we move to the study of optimal transportation for more general costs.

3 Costs behaving differntly in long-short regimes

Let $c(x, y) = h(x-y)$ be our cost of tranportation. The usual results in the theory of optimal tranportation heavily depend on the the injectivity of the map $\nabla h : \mathbb{R}^n \to \mathbb{R}^n$ that guarentee the well-definedness of the expression

$$s(x) = x - (\nabla h)^{-1}(\nabla \psi(x))$$

So the theory in its face-value is unable to deal with costs that do not possess a convex-type geometry and hence do not have an injective ∇h. The question that drives us in this work is the following: Does the invertibility of ∇h play an essential role in the single-valuedness of the optimal transference plan, or was it just an accident, a result of the particular method of proof employed above that the truth of Monge property seemed to be linked to invertibility of ∇h.

So we ask what could be said about Monge optimal tranportation for more general costs. Of course, as we have seen from one of our examples in the introduction3 if the cost possesses local minimas away from zero Monge problem wouldn't have any solution in general by the same type of couterexample built there. However, we conjecture the following:

Conjecture 1 *Let h be a positive strictly increasing C^1 function with $h(0) = 0$, $h(x)$ strictly convex near the origin for $x < r$, and strictly conave in long range $x > r$. Then there exist an optimal map $s : \mathbb{R}^n \to \mathbb{R}^n$ solving the optimal transport problem for $c(x, y) = h(|x - y|)$ in $\Pi(\mu, \nu)$ if we assume μ and ν are absolutely continuous with respect to Lebesgue and have disjoint compact supports.*

Remark. Let's see how far the usual geometric theory of optimal transportation takes us for this question. Since the cost is continuous, there would still exist a c-concave potential φ such that $\partial^c\varphi$ contains the support of all optimal Kantorovitch transference plans,

$$\forall\, \pi \in \Pi(\mu, \nu) \quad \text{c-optimal} \Rightarrow \text{Supp}(\pi) \subset \partial^c(\varphi)$$

What fails in the usual theory is that for many $\nabla\varphi(x)$ there would be usually two different y_1 and y_2 solving,

$$\nabla\varphi(x) = \nabla h(x - y_1) = \nabla h(x - y_2)$$

This undeterminacy of value of y is because ∇h repeats similiar values in short and long ranges. So when this occurs we have $|x - y_1| < r$ and $|x - y_2| > r$. Now this undeterminacy whould have been trivially avoided in cases when only one of y_1 and y_2 are in support of taget measure ν. But when there are points $x \in \text{Supp}(\mu)$ such that both y_1 and y_2 are in support of ν then the above Monge question becomes interesting.

It seems that from point of view of economics the above class of costs might be quite natural as we might expect that tranportation of goods in a local neighborhood to have a different fine behavior compared to transportation in long ranges. This is because we expect for transportation in different ranges, different means of transportation would be utilized. Since the means of transportation operating in different ranges are expected to possess different cost structures, studying these type of costs seem natural to us.

3.1 The Simplest Interesting Cost

Despite our interest in the main conjecture, there is however a more elementary cost function that we first want to look at. This cost is probably most basic example of cost behaving differently in short-long range. A positive resolution of Monge problem for this cost raises our confidence in the truth of main conjecture.

Although our interest in this cost was initially motivated by our interest in the main conjecture, it turned out that this cost is quite interesting in its own right. Indeed through studying this cost we were led to consider a new metric on space of compactly supported measures with very interesting properties. We'll explain these connections in the next part of this section.

Define

$$c_{01}(x, y) = \begin{cases} 1 & |x - y| > 1 \\ 0 & |x - y| \leq 1 \end{cases}$$

We shall call this cost, the 01 cost and denote it by c_{01}. This cost has transition point at distance 1 where the shift from short to long range beahvior occurs. Also one can imagine that this cost can easily arise as the limit of a sequence of costs that get more and more convex in short range, and more and more concave in long range. Also since c_{01} is lower-semicontinuous, Kantorovitch problem has a solution. However, one major difficulty with this cost is that it is not continuous. Now there are two places that continuity of cost played a major role in the general theory of optimal transportation:

1. In the introduction, When we proved that infimum cost for the Kantovitch optimal tranportation problem is the same as infimum cost for the Monge problem 1.

2. In the geometric theory of optimal transportation: to prove that the support of optimal Kantorovitch problem is c-cyclically monotone we needed the continuity.

The first problem is important because if we have any hope of being able to prove Monge problem via showing that a particularly chosen Kantovitch's solution is single-valued, as we did for the case of strictly convex cost, we must resolve the first difficulty.

The second problem seems unavoidable, hence c-cyclical monotonicity approach might be useless in this case. Fortunately, we have another tool that can be helpful in proving Monge problem for this cost, namely the Kantorovitch duality approach and theorem 6.

So our proposed approach to solve Monge problem for this cost is to do a careful analysis and chracterization of conjugate c-concave pairs for c_{01}. The author's experiments in dimension 1 shows that a remarkably interesting structure and chracterization must exist for cojugate c-concave pairs of this cost. In a sequel to this work we hope to be able to prove the next conjecture, Monge property for above cost, at least dimension 1 by using theorem 6.

Conjecture 2 *By $c_{01}(x, y)$ defined as above, there exist a Kantorivitch solution $\pi \in \Pi(\mu, \nu)$ which is almost everywhere single-valued, i.e. it solves the Monge problem .*

$$\pi = (Id \times T)_{\#}\mu$$

As explained above, our proposed stategy for above conjecture is to first solve it in dimension 1 via the theorem 6 and then use a similiar approach along the lines of work of Cafferelli and et al [3, 5] in $\mathcal{L}^1$ theory of optimal tranportation to reduce the case of higher dimension to one dimensional case.

Now we should go back to our first concern. The theorem below reassures us that at least in dimension 1, the Monge infimum cost coincides with Kantorovitch's infimal cost. Note that if above conjecture is true, the infimal Monge is equal to infimal Kantorovitch problem in all dimensions.

Proposition 2 *Assume μ and ν regular Borel probability measures over $\mathbb{R}$ and μ is absolutely continuous over $\mathbb{R}$. For any plan $\gamma \in \Pi(\mu, \nu)$ and $\epsilon > 0$, there exist a mapping T with cost less than $\mathcal{C}_{01}(\gamma) + \epsilon$ pushing μ to ν.*

Proof of above proposition is relegated to the appendix. The following lemma proves the conjecture 2 in dimension 1 for the special case when the cost of transportation between μ and ν is zero:

Lemma 3 *Let γ be a finite measure on $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν and assume μ is absolutely continuous with respect to Lebesgue measure. If cost of γ is zero, i.e.*

$$(x, y) \in Supp(\gamma) \quad \Rightarrow \quad |x - y| \leq 1$$

then there exist an optimal map, i.e. zero cost, pushing μ to ν.

Proof. We claim the canonical monotone push forward map $s(x)$ given by

$$\int_{-\infty}^{x} d\mu = \int_{-\infty}^{s(x)} d\nu$$

has zero cost. By restriction on support of γ we have:

$$\mu[(-\infty, x)] = \gamma[(-\infty, x) \times (-\infty, x+1)] \leq \nu[-\infty, x+1]$$
$$\nu[(-\infty, x)] = \gamma[(-\infty, x+1) \times (-\infty, x)] \leq \mu[-\infty, x+1]$$

which proves $|s(x) - x| \leq 1$ which means $s(x)$ is zero cost. Note that$s(x)$ is increasing hence Borel measurable. $\square$

Remark. The canonical monotone map in $\mathbb{R}^1$ does not in general solve the Monge problem for our cost, it only does it generally when the cost is zero. For example, let μ be lebesgue measure restricted to $[0, 1]$ and ν be the Lebesgure measure restricted to $[1.5, 2.5]$. Then the canonical monotone map induces a cost 1 map which is not infimal as one can easily construct a mapping with cost $\frac{1}{2}$ by mapping $[0.5, 1]$ to $[1.5, 2]$ via $+1$ translation which costs nothing and then take any map for the rest of the mass at the cost only $\frac{1}{2}$.

Above lemma motivates the following question: Given μ and ν when can we be sure that there exist a zero cost transference plan between them? Following proposition answers this question.

Proposition 3 ***(Characterization Property)****For any Borel set $E \subset \mathbb{R}^n$, let E^1 denote the set of points with distance one or less to some point of E. Let μ and ν be compactly supported probability measures on $\mathbb{R}^n$ absolutely continuous with respect to Lebesgue. Then for $\mathcal{C}_{01}(\mu,\nu)=0$ is necessary and sufficient that we have,*

$$\nu(E^1) \geq \mu(E) \qquad \forall\ E \subset \mathcal{B}_{\mathbb{R}^n}$$

Remark. The necessity is of course very easy. Because if $\pi \in \Pi(\mu,\nu)$ has a c_{01} zero cost, we can write

$$\mu(E) = \pi[E \times \mathbb{R}^n] = \pi[E \times E^1] \leq \pi[\mathbb{R}^n \times E^1] = \nu[E^1]$$

Here instead of assuming that the condition holds for every $E \in \mathcal{B}_{\mathbb{R}^n}$, we could have restricted the condition to open sets E. Actually if the condition

$$\nu(E^1) \geq \mu(E) \qquad \forall\ E \subset \mathbb{R}^n \text{ open}$$

holds then

$$\nu(F^1) \geq \mu(F) \qquad \forall\ F \subset \mathbb{R}^n \text{ Borel}$$

We shall prove this and the fact that E^1 is Borel measurable whenever E is Borel measurable in the appendix.

Because of the approximation procedure we use in the proof of above proposition, it was necessary to assume μ and ν were compactly supported and absolutely continuous with respect to Lebesgue. However, it seems almost certain that these assumptions can be dropped with a more sophisticated arguement. Our approximation procedure reduces our problem to the finite case. The finite case of the above theorem is actually the well-known *Hall's Marriage Theorem.* As a consequence of Hall's theorem, the proposition immeditely follows for the case when μ and ν are sum of the dirac deltas with equal weights.

Theorem 8 ***(Hall's Marriage Theorem)*** *Let $G = (S+T, E)$ be a bipariate graph with $2n$ vertices partioned into two equal sides S and T. For any $R \subset S$ define $N(R) \subset T$ as the set of neighboring vertices to R. So, $t \in N(R)$ if and only if there an edge in E connecting t to some $r \in R$. Then if there exist a perfect matching between S and T, if and only if*

$$|N(R)| \geq |R| \qquad \forall R \subset S$$

A perfect matching is a set of n edges in E connecting each element of S to a unique element of T.

Now assume $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{p_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{q_i}$. We build a graph G for our transportation problem. Associate to each p_i a vertex s_i in S and to each q_i a vertex t_i in T. [8] Connect s_i to t_j by an edge e_{ij} if $|p_i - q_j| \leq 1$ in $\mathbb{R}^n$. Now assume the condition of the proposition.

$$\nu(F^1) \geq \mu(F) \qquad \forall\ F \subset \mathbb{R}^n \text{ Borel}$$

by taking F to be the union of a subset of $\{p_1, p_2, p_3, \ldots, p_n\}$, the conditon translates to the condition of Hall's Marriage theorem. As a consequence there exist a perfect matching

$$\{\ (\ i, \sigma(i)\ )\ \}_{i=1}^{n}\ \subset\ E \quad \Longrightarrow \quad |p_i - q_{\sigma(i)}| \leq 1$$

Hence

$$\pi = \sum_{i=1}^{n} \delta(x - p_i)\delta(y - q_{\sigma(i)})\ \in\ \Pi(\mu, \nu)$$

solves the Monge problem.

The idea of the proof of *Chracterization Property Proposition* is to approximate μ and ν by sequences $\{\mu_k\}_{k=1}^{\infty}$ and $\{\nu_k\}{-k} = 1^{\infty}$ such that μ_k's and ν_k are sums of dirac deltas of equal weight, and so Hall's theorem applies to them. Our approximation must be such that the condition

$$\nu(F^1) \geq \mu(F) \qquad \forall\ F \subset \mathbb{R}^n \text{ Borel}$$

translates approximately to Hall's condition for (μ_k, ν_k) pairs. Hence we can find $\pi_k \in \Pi(\mu_k, \nu_k)$ solving this the approximate (μ_k, ν_k) problem with zero cost. Then by applying compactness coming from Prokhorov's theorem to $\{\pi_k\}_{k=1}^{\infty}$, we can find $\pi \in \Pi(\mu, \nu)$ the limit point of the sequence $\{\pi_k\}_{k=1}^{\infty}$ which would be concenterated on the set

$$\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid |x - y| \leq 1\}$$

and hence would be our zero cost solution to the original (μ, ν) problem.

The detailed proof of proposition can be found in the appendix.

[8] Here it is possible that two points with different indices p_i, p_j are actually situated at the same location in $\mathbb{R}^n$. Even when that happens p_i and p_j are associated with two different vertices v_i and v_j. Of course in that case v_i and v_j will have exactly the same neighboring vertices in T.

3.2 A New Metric on $\mathcal{P}_c(\mathbb{R}^n)$

In previous part, we considered a cost c_{01} with different behavior in short range versus long range, and we answered the chracterization question for when there exist a zero cost transference plan between μ and ν. For this cost the transition from short to long range behavior occurred at distance 1, however, we could have actually considered a family of costs parametrized by a variable $s > 0$ that signifies the distance at which the transition form short to long range occurs.

$$c_{01}^s(x,y) = \begin{cases} 1 & |x-y| > s \\ 0 & |x-y| \leq s \end{cases}$$

Let $\mathcal{C}_s$ denote the optimal transportation cost for c_{01}^s.
Let μ and ν be compactly supported probabality measures on $\mathbb{R}^n$.[9] For large enough $s > \mathrm{diam}(\mathrm{Supp}(\mu) \bigcup \mathrm{Supp}(\nu))$, $\mathcal{C}_s(\mu,\nu) = 0$. On the other hand, by applying chracterization property to c_{01}^s cost

$$\mathcal{C}_s(\mu,\nu) = 0 \quad \Longleftrightarrow \quad \nu(F^s) \geq \mu(F) \quad \forall F \in \mathcal{B}_{\mathbb{R}^n}$$

As $s \to 0$ if the second condition $\nu(F^s) \geq \mu(F)$ and its equaivalent condition $\mu(F^s) \geq \nu(F)$ hold for all F, [see Lemma 4], it means that $\mu = \nu$. So if $\mu \neq \nu$, we have $\mathcal{C}_s(\mu,\nu) > 0$ for some $s > 0$. We are interested in the point of transition, i.e.

$$\mu,\nu \in \mathcal{P}_c(\mathbb{R}^n): \quad R(\mu,\nu) := \inf_{s>0} s \qquad \mathcal{C}_s(\mu,\nu) = 0$$

. It turns out that this number defines a new metric on the space of compactly supported probability measures $\mathcal{P}_c(\mathbb{R}^n)$. Now we try to write this definition more concretely:

Definition 6 *Fix $\mu,\nu \in \mathcal{P}_c(\mathbb{R}^n)$. For any $\pi \in \Pi(\mu,\nu)$, cosider*

$$I^\infty(\pi) := \sup_{(x,y)} |x-y| \qquad (x,y) \in Supp(\pi)$$

Then define

$$W_\infty(\mu,\nu) := \inf_\pi I^\infty(\pi) \qquad \pi \in \Pi(\mu,\nu)$$

[9] Most of the definition and theorems in this section can be done for general Polish spaces but for the sake of concreteness and also because we are ultimately interesed in $\mathbb{R}^n$, we restric ourselves only to this case.

Proposition 4 *The two defnitions of W_∞ are equaivalent, i.e.*

$$W_\infty(\mu,\nu) = \inf_{\mathcal{C}_s(\mu,\nu)=0} s$$

Moreover, there exist some $\pi \in \Pi(\mu,\nu)$ such that

$$I^\infty(\pi) = W_\infty(\mu,\nu)$$

Proof. Let $\epsilon > 0$. Let $r = W_\infty(\mu,\nu) + \epsilon$. Take $\pi_r \in \Pi(\mu,\nu)$ such that $I^\infty(\pi) < r$. $\mathcal{C}_r(\mu,\nu) = 0$ by existence of π_r. Hence as $r \to 0$, we get

$$W_\infty(\mu,\nu) \geq \inf_{\mathcal{C}_s(\mu,\nu)=0} s$$

The other side of the inequality is equally clear.

Now, notice that $I^\infty : \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n) \to \mathbb{R} \cup \{\infty\}$ is lower-semicontinuous in weak topology. If $\{\gamma_k\}_{k=1}^\infty$ is a sequence that weakly converges to γ in $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$, we know for any $(x,y) \in \mathrm{Supp}(\gamma)$ there must exist a sequence $\{(x_k,y_k)\}_{k=1}^\infty$ that converges to (x,y) as $k \to \infty$. Hence

$$I^\infty(\gamma) \leq \liminf_{k\to\infty} I^\infty(\gamma_k)$$

Since I^∞ is l.s.c. and $\Pi(\mu,\nu)$ is compact by Prokhorov's theorem, we can always find a minimizer π.

Alternatively, we could have arrived at a minimizing π by taking a sequence $r_k \downarrow W_\infty(\mu,\nu)$ and taking the minimizers π_{r_k} solving the kantorovitch problem for $c_{01}^{r_k}$. By compactness of $\Pi(\mu,\nu)$, this sequence would have a limit point which satisfies our condition. $\square$

Next thing we need to show W_∞ is a metric. The fact that

$$W_\infty(\mu,\nu) = 0 \iff \mu = \nu$$

and that W_∞ is symmetric are clear. The only thing is to show triangle inequality. But this again is a very easy consequence of the gluing lemma2. Let μ_1, μ_2 and μ_3 be three probability measures in $\mathcal{P}_c(\mathbb{R}^n)$. Let π_{12} and π_{23} respectively be the optimal measures for $W_\infty(\mu_1,\mu_2)$ and $W_\infty(\mu_2,\mu_3)$ and let π be the gluing of π_{12} and π_{23}. Then since marginals of π are π_{12} and π_{23}, we have

$$\forall\ (x,y,z) \in \mathrm{Supp}(\pi) \ \Rightarrow\ (x,y) \in \mathrm{Supp}(\pi_{12}) \quad (y,z) \in \mathrm{Supp}(\pi_{23})$$

Hence if you take π_{13} the marginal of π given by reflecting on 1 and 3 factor, for any (x,z) a point in $\mathrm{Supp}(\pi_{13})$,

$$|x-z| \leq |x-y| + |y-z| \leq W_\infty(\mu_1,\mu_2) + W_\infty(\mu_2,\mu_3)$$

Hence we have our triangle inequality

$$W_\infty(\mu_1,\mu_2) \leq I^\infty(\pi) \leq W_\infty(\mu_1,\mu_2) + W_\infty(\mu_2,\mu_3)$$

So the next question after this would be what topology does this new metric W_∞ induces on $\mathcal{P}_c(\mathbb{R}^n)$.

Proposition 5 *Let $\{\mu_k\}_{k=1}^\infty$ be a sequence of probability measures such that $W_\infty(\mu_k,\mu) \to 0$. This condition implies the weak convergence of μ_k to μ with the W_∞ tightness condition*

$$\exists\, K \subset \mathbb{R}^n \text{ compact } \Rightarrow \mathit{Supp}(\mu_k) \subset K \quad \forall k \in \mathbf{N}.$$

To prove this, it is most convinient for us to wait till after we prove the fundamental theorem about W_∞ that links this metric with the Monge-Kantorovitch distances W_p. So the choice of notation for W_∞ wasn't a coincidence. Knowing above proposition such link between W_p , $p > 1$ and W_∞ is not that surprising because as we mentioned W_p also metrizes the weak convergence modulu a p-th moment tightness condition. Indeed, W_∞ tightness condition implies all p-th moment tightness conditions, and so above proposition shows if $W_\infty(\mu_k,\mu) \to 0$, then for all $p > 1$, $W_p(\mu_k,\mu) \to 0$ as well.

However, something much stronger is true which is the content of the fundamental link between W_∞ and W_p's:

Proposition 6 *Fix μ and ν compactly supported probability measures on $\mathbb{R}^n$. Recall the notation*

$$I_p(\pi) = \left(\int_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^p \, d\pi(x,y)\right)^{\frac{1}{p}}$$

and the definition of $W_p = I_p(\pi_p)$, where π_p denotes one of the minimizers of p-th Monge Kantorovitch distance. We have,

$$W_\infty(\mu,\nu) = \lim_{p\to\infty} W_p(\mu,\nu)$$

and if $\pi_{p_k} \to \pi$ weakly as $p_k \to \infty$ then

$$I^\infty(\pi) = W_\infty(\mu,\nu)$$

which means π is a minimizer of W_∞ distance.

Note that $\lim_{p\to\infty} W_p(\mu,\nu)$ always makes sense because W_p is increasing with respect to p.

Proof.

First of all we have the measure theoretic result that for any γ probability measure on $\mathbb{R}^n$

$$\lim_{p\to\infty} I_p(\gamma) = \sup_p I_p(\gamma) = I^\infty(\gamma)$$

The first equality is true because again by Holder $I_p(\gamma)$ is increasing with respect to p. Moreover since $|x-y| \leq I^\infty(\gamma)$ for any (x,y) in support of γ,

$$I_p(\gamma) = \left(\int_{\mathbb{R}^n\times\mathbb{R}^n} |x-y|^p \, d\gamma(x,y)\right)^{\frac{1}{p}} \leq I^\infty(\gamma)$$

To prove the otherside of inequlity, we will show for any $\epsilon > 0$,

$$\lim_{p\to\infty} I(p)(\gamma) > I^\infty(\gamma) - \epsilon$$

Let $r = I^\infty(\gamma) - \epsilon$. There exist (x,y) in support of γ such that $|x-y| > r + \frac{\epsilon}{2}$. Let B be a ball of radius less than $\frac{\epsilon}{4}$ around (x,y). Since $\gamma(B) > 0$ and $|x'-y'| > r$ for all (x',y') in B, we have,

$$I_p(\gamma) \geq \gamma(B)^{\frac{1}{p}} \cdot r$$

As we let $p \to \infty$, we get the other side of inequlity. Hence,

$$\lim_{p\to\infty} I_p(\gamma) = \sup_p I_p(\gamma) = I^\infty(\gamma)$$

Now we know,

$$W_p(\mu,\nu) = I_p(\pi_p) \leq I_p(\pi_\infty)$$

Letting p to infinity we get

$$\limsup_{p\to\infty} I_p(\pi_p) \leq I_\infty(\pi_\infty) \qquad \text{(eqn 1)}$$

Now Consider a sequence $\{\pi_{p_i}\}_{i=1}^\infty$ such that $p_i \to \infty$. By compactness of $\Pi(\mu,\nu)$ and passing to a subsequence we can assume, $\pi_{p_i} \to \tilde{\pi}$. Of course,

$$\tilde{\pi} \in \Pi(\mu,\nu)$$

We know

$$I^{\infty}(\tilde{\pi}) \geq I^{\infty}(\pi_{\infty}) \qquad \text{(eqn 2)}$$

So there exists a $q_0 = (x_0, y_0)$ in Support of $\tilde{\pi}$ such that

$$|x_0 - y_0| \geq I^{\infty}(\pi_{\infty}) - \frac{\epsilon}{2}$$

So there exist a ball B' with $\tilde{\pi}(B') > 0$,

$$B' \subset \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid \ |x - y| > I^{\infty}(\pi_{\infty}) - \epsilon\}$$

Because $\tilde{\pi}$ is the weak limit of π_{p_i},

$$\pi_{p_i}(B') \to \tilde{\pi}(B') \qquad \text{as } \ i \to \infty$$

But then

$$I_{p_i}(\pi_{p_i}) \geq |\, I^{\infty}(\pi_{\infty}) - \epsilon \,| \cdot (\pi_{p_i}(B'))^{\frac{1}{p_i}}$$

Take N large enough that for all $i > N$, $\pi_{p_i}(B') > k$ for some $0 < k \leq \tilde{\pi}(B')$. Then for $i > N$, $(\pi_{p_i}(B'))^{\frac{1}{p}} \geq k^{\frac{1}{p_i}}$ and as $i \to \infty$ and $p_i \to \infty$, we get $k^{\frac{1}{p_i}} \to 1$.
Therefore

$$\liminf_{p_i \to \infty} I_{p_i}(\pi_{p_i}) \geq I^{\infty}(\pi_{\infty}) - \epsilon$$

holds for every small $\epsilon > 0$ and so,

$$\liminf_{p_i \to \infty} I_{p_i}(\pi_{p_i}) \geq I^{\infty}(\pi_{\infty}) \qquad \text{(eqn 3)}$$

Now note that if in eqn 2 strict inequality would have hold, i.e. for some $\delta > 0$

$$I^{\infty}(\tilde{\pi}) > I^{\infty}(\pi_{\infty}) + \delta$$

We could have used this analogously to prove,

$$\liminf_{p_i \to \infty} I_{p_i}(\pi_{p_i}) \geq I^{\infty}(\pi_{\infty}) + \delta$$

But this is impossible because we have already shown eqn 1

$$\limsup_{p \to \infty} \ I_p(\pi_p) \leq I_{\infty}(\pi_{\infty})$$

So we conclude that $\tilde{\pi}$ weak limit point of $\{\pi_{p_k}\}_{k=1}^{\infty}$ as $p_i \to \infty$

$$I^{\infty}(\tilde{\pi}) = I^{\infty}(\pi_{\infty})$$

Also combining eqn 1 and eqn 3 we get our desired result

$$W_{\infty}(\mu, \nu) = \lim_{p\to\infty} W_p(\mu, \nu) \quad \square$$

Now let's quickly prove the proposition 5,

Proof of Proposition 5. Assume $W_{\infty}(\mu_k, \mu) \to 0$. Since μ_k's are compactly supported, μ must be compactly supported otherwise all the terms $W_{\infty}(\mu_k, \mu)$ are infinite. Now take $r = \sup_k W_{\infty}(\mu_k, \mu) < +\infty$. Then all μ_k's must be supported in at most distance r of $\mathrm{Supp}(\mu)$. Now we prove μ_k converge weakly to μ. Using previous proposition, this is easy because[10]

$$W_{\infty}(\mu_k, \mu) \to 0 \quad \Rightarrow \quad W_p(\mu_k, \mu) \to 0$$

But we already know $W_p(\mu_k, \mu) \to 0$ implies weak convergence of μ_k to μ.[1] $\square$

Example 4 *The other direction of this proposition is not true. If $\{\mu_k\}_{k=1}^{\infty}$ converges weakly to μ and all μ_k's have a commomn support, it doesn't necessarily mean that $W_{\infty}(\mu_k, \mu) \to 0$ as k goes to infinity. Let $\mu = \mathcal{L}^1[0, 1]$ and $\mu_k = \frac{k-2}{k}\mathcal{L}^1[0, 1] + \frac{1}{k}\delta(x - 2) + \frac{1}{k}\delta(x + 1)$. This is a counterexample.*

Example 5 *One might suspect because in euclidean case π_p minimizing $W_p(\mu, \nu)$ was unique, the limiting measure minimizing $W_{\infty}(\mu, \nu)$ might be unqiue as well. This is not true as the set π_p might not be cauchy and hence it might have many limit points. Indeed this occurs: If you take $\mu = \frac{1}{3}(\delta(x) + \delta(x - 2) + \delta(x - 50))$ and $\nu = \frac{1}{3}(\delta(x - 1) + \delta(x - 3) + \delta(x - 100))$, there are many $\pi \in \Pi(\mu, \nu)$ having the minimal $I^{\infty}(\pi) = 50$.*

There is one property that is preserved as we go down the limit from W_p to W_{∞} and that is the *displacement interpolation property*. By contracting along the geodesic path associated to each π_p , we created a family π_p^s interpolating between $(Id \times Id)_{\#\mu}$ and π_p. Pick a sequence $p_k \to \infty$. We can assume $\pi_{p_k} \to \pi$ for some $\pi \in \Pi(\mu, \nu)$.

[10] This could have been proved with a bit more effort without previous theorem.So this is not that essential of an application.

Now consider the interpolated families. Since the interpolation just contract the domain of measures along the diagonal, it preseveres the weak limit, therefore

$$\pi^s_{p_k} \to \pi^s \qquad \forall\, 0 \leq s \leq 1$$

Since $\pi^s_{p_k}$ are optimal for each W_{p_k} by the main proposition 6, whatever the limit point of them is, must be optimal for W_∞ as well. So π_s the interpolated family for π is optimal. Moreover,

$$I^\infty(\pi_s) = \lim_{k\to\infty} I_{p_k}(\pi^s_{p_k}) = s \cdot \lim_{k\to\infty} I_{p_k}(\pi_{p_k}) = s \cdot I^\infty(\pi)$$

So the cost also changes linearly. (Although this fact is actually much more clear from the picture of $\mathrm{Supp}(\pi^s)$ family.)

Definition 7 *Let $\Pi_{opt}(\mu,\nu)$ stand for the set of all $\pi \in \Pi(\mu,\nu)$ such that*

$$I^\infty(\pi) = W_\infty(\mu,\nu)$$

And similiarly let $\Pi^{\lim}_{opt}(\mu,\nu)$ denote the distinguished subset of $\Pi_{opt}(\mu,\nu)$ which can be obtained from the limit of $\{\pi_{p_k}\}$'s as proposition 6 suggests.

In this language we have proved the following corollary:

Corollary 2 *For $\pi \in \Pi^{\lim}_{opt}(\mu,\nu)$ the displacement interpolation property holds which means all members of the family π^s are W_∞-optimal.*

Now this corollay looks a little impotent in the first sight. But it actually must be sufficient in most applications because we can always choose to pick $\pi \in \Pi^{\lim}_{\mathrm{opt}}(\mu,\nu)$ because thanks to proposition 6 we know $\Pi^{\lim}_{\mathrm{opt}}(\mu,\nu)$ is non-empty. However, it remains an interesting question whether displacement interpolation property holds for all elements of $\Pi_{\mathrm{opt}}(\mu,\nu)$ or not. Also it's generally important to understand what chracterizes the elements of $\Pi_{\mathrm{opt}}(\mu,\nu)$ that can be obtained via this limit.

However one thing that is for certain is,

$$\Pi_{\mathrm{opt}}(\mu,\nu) \neq \Pi^{\lim}_{\mathrm{opt}}(\mu,\nu)$$

For example, in dimension 1 all π_p's for $p > 1$ are monotone, hence the weak limit is also monotone. But as one can easily construct examples by similiar construction to example5, there can be many non-monotone W_∞ optimizers.

References

[1] Villiani, Cedric. Topics in Optimal Tranportation. *Graduate Studies in Mathematics Volume 58.* (2003)

[2] Gangbo, Wilfrid and Mccan, R.J. The Geometry of Optimal Transportation. *Acta Math, 177* (1996) p113-161.

[3] Ambrosio, L. and Pratelli A. Existence and stability results in $\mathcal{L}^1$ theory of Optimal Transportation. *Ambrosio Lecture notes* (2003)

[4] Villiani, Cedric. Optimal Transport, old and new. *Grundlehren der mathematischen Wissenschaften, Volume 338.* Springer(2009).

[5] Caffarelli, L. A., Feldman M. and Mccan R.J. Constructing Optimal Maps For Monge's Transport Problem As a Limit of Strictly Convex Costs. *Journal Of American Mathematical Society.* Volume 15, Number 1, p1-26. (2001)

A Appendix

Proof of Proposition2. First assume γ gives zero mass to both lines $x = y+1$ and $x = y-1$ which means that no mass is transfered via the point of discontinutiy of cost at $|x-y| = 1$. Let us we donote by $D = L_1 \cup L_2$ the union of these two lines,

$$\gamma[D] = 0$$

Now define $D_n = \{(x, y) \in \mathbb{R}^2, 1 - \frac{1}{n} < |x-y| \leq 1\}$ and define the cost $c_n(x, y) = E_n(|x-y|)$ by,

$$E_n(x) = \begin{cases} 0 & x \leq 1 - \frac{1}{n} \\ n(x - 1 + \frac{1}{n}) & 1 - \frac{1}{n} < x \leq 1 \\ 1 & x > 1 \end{cases}$$

So $c \leq c_n$ and c_n only differs with c over D_n. Since the intersection of D_n's are D which is measure zero, you can find n such that $\gamma[D_n] < \frac{\epsilon}{2}$. Then we have,

$$\mathcal{C}_n[\gamma] < \frac{\epsilon}{2} + \mathcal{C}(\gamma)$$

Also by continuity of c_n there exist mapping s only at most $\frac{\epsilon}{2}$ more expensive than γ, we get the following,

$$\mathcal{C}(s) \leq \mathcal{C}_n(s) \leq \mathcal{C}_n(\gamma) + \frac{\epsilon}{2} \leq \mathcal{C}(\gamma) + \epsilon$$

So we are done in this case. Now given a γ that gives a positive measure to D, the strategy is to produce $\tilde{\gamma}$ still pushing μ to ν, but giving zero measure to D, while only increasing cost by $\frac{\epsilon}{2}$. After that, applying above result to $\tilde{\gamma}$ finishes the proof.

So assume $\gamma[D] > 0$. Without loss of generality then $\gamma[L_1] > 0$. So let γ_{L_1} be the restriction of γ to L_1 and let μ_1 be the projection of γ_{L_1} to first variable. Since μ is absolutely continuous with respect to Lebesgue measure then so is μ_1, and hence by Radon-Nikodym we can write $\mu_1 = f_1 \, dm$. Since γ_{L_1} is supported on $y = x+1$ line, we see that the projection of γ_{L_1} to the second variable must be just $+1$ tranlation of μ_1.

$$\nu_1 = f_1(y-1) \, dm$$

Now γ_{L_1} has zero cost and we want to find a new map $\tilde{\gamma}_{L_1}$ to push $f_1(x) \, dm$ to $f_1(x-1) \, dm$ without any mass on D and negiligble cost. The key fact that we shall use is that

$$\int_{\mathbb{R}} |f_1(x-1) - f_1(x-h)| \, dm \;\to\; 0 \qquad \text{as} \quad h \;\to\; 1$$

So we choose $h < 1$ close to 1 such that the above integral is as small as we please. So let $g(x)$ be the common mass between $+1$ and $+h$ push f_1, i.e.

$$g(x) := f_1(x-1) \wedge f_1(x-h)$$

So we have $g(x+h) \leq f(x)$ and $g(x) \leq f(x-1)$. Now we are ready to build our mapping $\tilde{\gamma}_{L_1}$. Consider

$$\begin{aligned} \pi : \mathbb{R} \;&\to\; \mathbb{R}^2 \\ x \;&\to\; (x, x+h) \end{aligned}$$

The first part of $\tilde{\gamma}_{L_1}$ is $\pi_{\#} \, g(x)$ which has zero cost since $0 < h < 1$. The leftover is $l(x) = f_1(x) - g(x+h)$ that must be pushed to $l'(x) = f_1(x-1) - g(x)$. The leftover has a negligible mass by choice of h, so we just need to find any map between them wit zero concenteration on D. This is easy just take the map $r(x)$ to be the canonical orientation reversing map

$$\int_{-\infty}^{x} l(x) \, dm = \int_{r(x)}^{\infty} l'(x) dm$$

Since as x increases $r(x)$ decreses, it is at only two points that $|r(x)-x| = 1$ happens. Hence the map has zero measure on D. So we produce:

$$\tilde{\gamma}_{L_1} = \pi_{\#}\, g(x) + (Id \times r(x))_{\#}\, l(x)$$

By similiar construction we can do the job as well for L_2. Our desired $\tilde{\gamma}$ is then given by,

$$\tilde{\gamma} = \gamma - \gamma_{L_1} - \gamma_{L_2} + \tilde{\gamma}_{L_1} + \tilde{\gamma}_{L_2}$$

and so we are done. □

Here we present a Lemma related to the remakrs followed by the chracterization property proposition3.

Lemma 4 *Assume μ and ν are Borel probability measures on $\mathbb{R}^n$. The condition*

$$\nu(F^1) \geq \mu(F) \qquad \forall\, F \subset \mathbb{R}^n \text{ open}$$

is equivalent with

$$(i)\ \nu(E^1) \geq \mu(E) \qquad \forall\, E \subset \mathbb{R}^n \text{ Borel}$$

and also

$$(ii)\ \mu(E^1) \geq \nu(E) \qquad \forall\, E \subset \mathbb{R}^n \text{ Borel}$$

Moreover E^1 is always Borel measurable for any set E, so above expressions make sense.

The first equivalence says that we could have instead worked with open sets and the second condition says that the role of μ and ν is symmetric.

Proof. For any set we defined

$$E^s := \{p \in \mathbb{R}^n \mid \inf_{q \in E} |p - q| \leq s\}$$

Now we define

$$E^s_+ := \{p \in \mathbb{R}^n \mid \inf_{q \in E} |p - q| < s\}$$

Note that for any set E, E^s_+ is always an open set. Let $r_i > 0$ be a sequence such that $r_i \downarrow 0$. Notice that

$$E^1 = \bigcap_{i=1}^{\infty} E_+^{1+r_i}$$

So E^1 is Borel measurable. Similiary closure of E denoted by $\overline{E}$ can be written as $\bigcap_{i=1}^{\infty} E^{r_i}$. We have,

$$\nu(E^1) = \inf_{k\to\infty} \nu(E^{1+r_k}) \geq \inf_{k\to\infty} \mu(E^{r_k}) = \mu(\overline{E}) \geq \mu(E)$$

Which proves the (i). For (ii), let $F = (E^1)^c$. Then $F^1 = \big((E^1)^c\big)^1 \subset E^c$. So

$$\mu(E^1) \;\geq\; \nu(E) \quad\iff\quad \nu(E^c) \;\geq\; \mu\big(\big(E^1\big)^c\big) = \mu(F)$$

But then from the hypothesis $\nu(F^1) \geq \mu(F)$. it follows that

$$\nu(E^c) \geq \nu(F^1) \geq \mu(F) \implies \mu(E^1) \;\geq\; \nu(E) \quad \square$$

Now we should present the proof of proposition3, the chracterization property. Notice that we will use the assumption of absolute continuity and compact support in the proof. However, we believe these assumptions are not essential and we hope that with a little more work they could be dropped.

Proof of Proposition3. We have already mentioned that to prove this, our plan is to approximate μ and ν by sequences of $\{\mu_n\}_{n=1}^{\infty}$ and $\{\nu_n\}_{n=1}^{\infty}$.

* $\mu_n \to \mu$ and $\nu_n \to \nu$ weakly as $n \to \infty$.
* All $\{\mu_n\}_{n=1}^{\infty}$ and $\{\nu_n\}_{n=1}^{\infty}$ have a common compact support.
* μ_k's and ν_k's are sum of dirac deltas of specific weight $\frac{1}{M_k}$ for each k.

However as it turns out in our construction, ν_n's and μ_n's are not going to be probability measures but instead μ_n's will approximate μ from below and ν_n's will approximate ν from above. Therefore we will have

$$\nu_n(\mathbb{R}^n) \geq 1 \qquad \mu_n(\mathbb{R}^n) \leq 1 \qquad \forall n \in \mathbf{N}$$

What our approximation accomplishes beside $\mu_n \to \mu$ and $\nu_n \to \nu$ is that there exist a sequence $\epsilon_n \downarrow 0$ such that for all $E \in \mathcal{B}_{\mathbb{R}^n}$,

$$\nu_n(E^{\epsilon_n}) \geq \nu(E) \qquad \mu(E^{\epsilon_n}) \geq \mu_n(E) \qquad \forall\, n \in \mathbf{N} \qquad \text{(Approximation)}$$

Because of this property, for every $E \in \mathcal{B}_{\mathbb{R}}^n$ we will have

$$\nu_n(E^{1+2\epsilon_n}) \;\geq\; \nu(E^{1+\epsilon_n}) \;\geq\; \mu(E^{\epsilon_n}) \;\geq\; \mu_n(E)$$

Here we kept using the fact that $(R^a)^b \subset R^{a+b}$. Also the middle inequlity in above is by the hypothesis and the other two by our Approximation property.

The fact that $\nu_n(E^{1+2\epsilon_n}) \geq \mu_n(E)$ is exactly the type of condition we need to use the Hall's theorem. So indeed, we associate a bipartie graph G_n with the pair (μ_n, ν_n) with vertices (S, T) such that each elements of S is associted with a dirac delta point of μ_n, and each element of T is associated with a dirac delta point of ν_n.

$$\mu_n = \frac{1}{M_n} \sum_{i=1}^{K_n} \delta(x - p_i) \qquad \nu = \frac{1}{M_n} \sum_{i=1}^{K'_n} \delta(y - q_i)$$

Notice here we droped the index n for S and T and p_i and etc to avoid cumbersome notation. But of cousem all these graphs are different for different n's.

Now the edges of G_n are made up of those $e = \{s_i, t_j\}$ such that for the associated points p_i and q_j,

$$|p_i - q_j| \leq 1 + 2\epsilon_n$$

So by Approximation property, the Hall's condition is satisfied in G_n. However, there is one catch and that is the fact that μ_n and ν_n were not probability measures. So the number of verticies may not be equal, so the Hall's theorem does not directly apply here.

$$|S| = K_n \leq K'_n = |T| \qquad \text{not necessarily equality)}$$

However, fortunately a generalization of Hall's theorem still holds,

Theorem 9 ***(Hall's Generalized Marriage Theorem)*** *Let $G = (S + T, E)$ be a bipariate graph with S and T not necessarily equal size. Then G contains a matching of S into T if and only if*

$$|X| \leq |N(X)|$$

for every $X \subset S$.

The conditions of this theorem holds for our graph G_n of (μ_n, ν_n) so we find a matching from S to T. Hence we can form

$$\pi_n = \frac{1}{M_n} \sum_{i=1}^{K_n} \delta(x - p_i)\, \delta(y - q_{\sigma(i)})$$

Now π_n has marginal μ_n and $\tilde{\nu}_n$. $\tilde{\nu}_n$ might differe a little from ν because it might not contain all q_j's because of some of indices j do not come into the matching and are not of the form $\sigma(i)$.

$$\tilde{\nu}_n = \frac{a}{M_n}\sum_{i=1}^{K_n} \delta(y - q_{\sigma(i)})$$

Note that $\nu_n - \tilde{\nu_n}$ is a positive measure so its total variantion norm is

$$\nu_n(\mathbb{R}^n) - \tilde{\nu}(\mathbb{R}^n) = \nu_n(\mathbb{R}^n) - \pi_n(\mathbb{R}^n \times \mathbb{R}^n) = \nu_n(\mathbb{R}^n) - \mu_n(\mathbb{R}^n)$$

But the condition $\mu_n \to \mu$ and $\nu_n \to$ combined with the condition of common compact support means that $\nu_n(\mathbb{R}^n) \to \nu(\mathbb{R}^n) = 1$ as well as $\mu_n(\mathbb{R}^n) \to \mu(\mathbb{R}^n) = 1$. So above difference vanishes and therefore

$$\tilde{\nu}_n \to \nu \qquad n \to \infty$$

By common support condition $\{\pi_k\}_{k=1}^{\infty}$ is tight and hence compact, so we can assume $\pi_k \to \pi$. Now clearly $\pi \in \Pi(\mu,\nu)$. Let (x, y) be a point in support of π. Then there must be a sequence $(x_n, y_n) \in \mathrm{Supp}(\pi_n)$ such that

$$(x_n, y_n) \to (x, y) \qquad n \to \infty$$

Since $|x_n - y_n| \leq 1 + 2\epsilon_n$ and $\epsilon_n \downarrow 0$ as we get,

$$|x - y| \leq 1$$

And hence π solves our chracterization property proposition.$\square$

It remains to show the existence of our desired approximation. Let $\mu = f\,dm$ where dm is lebesgue measure on $\mathbb{R}^d$ and f is compactly supported nonegative Borel measurable function. For our convinience, we do this in the language of dimension 2 and rectangles, however, the argument works for any dimension.

Approximate f from below by simple functions f_n such that f_n increase to f pointwise,

$$f_n = \sum_{i=1}^{A_n} a_i \cdot \chi_{R_i}$$

Here R_i's are rectangles. Choose M_n very large at least such that $M_n \geq 2M_{n-1}$, and also large enough such that if you choose

$$\frac{k_i^2}{M_n} \leq a_i \cdot |R_i| \leq \frac{(k_i+1)^2}{M_n}$$

the difference is very small, like

$$\sum_{i=1}^{A_n} \left(a_i \cdot |R_i| - \frac{k_i^2}{M_n} \right) \leq \frac{1}{M_{n-1}}$$

And also large enough such that

$$\epsilon_n = \sup_i \frac{|diam(R_i)|}{k_i} \leq \frac{1}{2}\, \epsilon_{n-1} \qquad 1 \leq i \leq A_n$$

So we produced a new approximation to $\mu = f\, dm$.

$$\tilde{f}_n = \frac{1}{M_n \cdot |R_i|} \sum_{i=1}^{A_n} k_i^2\, \chi_{R_i}$$

To produce μ_n divide each R_i into k_i^2 smaller rectanles of diameter $\frac{|diam(R_i)|}{k_i}$ by dividing the length and width of R_i into k_i parts. Now put k_i^2 dirac delta measures of weight $\frac{1}{M_n}$ in the center of each k_i^2 small rectangle of R_i. The sum of all these dirac deltas is our desired μ_n.

$$R_i = \bigcup_{m=1}^{k_i^2} R_i^m \qquad \mu_n = \frac{1}{M_n} \sum_{i=1}^{A_n} \sum_{m=1}^{k_i^2} \delta(x - p_{i,m})$$

In above $p_{i,m}$ is the center of R_i^m, the m-th out of k_i^2 rectangle in R_i.

Now we want to show $\mu(E^{\epsilon_n}) \geq \mu_n(E)$ for every E. This is clear because $\mu_n(E)$ really counts the number of the dirac delta points in μ_n that fall into E.

$$\mu_n(E) = \frac{1}{M_n} \, |\{p_{i,m} \in E\}|$$

Now by definition of ϵ_n, if E intersect any of those small rectangles R_i^m, which it has to if it contains the dirac delta point $\frac{1}{M_n}\delta(x - p_{i,m})$ at the center of R_i^m, then E^{ϵ_n} contains the whole of that R_i^m rectangle. So

$$\mu(E^{\epsilon_n}) = \int_{E^{\epsilon_n}} f\,dm \geq \int_{E^{\epsilon_n}} \tilde{f}_n\,dm$$
$$\geq \sum_{p_{i,m}\in E} \frac{k_i^2}{M_n \cdot |R_i|}\,|R_i^m| = \sum_{p_{i,m}\in E} \frac{1}{M_n} = \mu_n(E)$$

By similiar methods we construct ν_n from ν, but for ν_n's we will use approximation from above instead. So we get an approximating sequence $\{\nu_n\}_{n=1}^{\infty}$ with $M'_n \uparrow \infty$ and $\epsilon'_n \downarrow 0$. Producing a common $\{\epsilon_n\}_{n=1}^{\infty}$ and $\{M_n\}_{n=1}^{\infty}$ is easy because we can always take $Max\{\epsilon_n, \epsilon'_n\}$ sequence, and we can always replace each dirac delta of weight $\frac{1}{M_n}$ by M'_n dirac deltas of weight $\frac{1}{M_n \cdot M'_n}$. By this adaptation, it is easy to see all conditions including the crucial Approximation would be kept.

The only non-trivial thing left to prove is that $\mu_k \to \mu$ as $k \to \infty$. Let $k \geq n$ then,

$$\mu(E^{\epsilon_n}) \;\geq\; \mu_k(E)$$

Let $k \to \infty$,

$$\mu(E^{\epsilon_n}) \;\geq\; \limsup_{k\to\infty} \mu_k(E) \qquad \forall n \in \mathbf{N}$$

Therefore,

$$\mu(E) \geq \limsup_{k\to\infty} \mu_k(E) \tag{Supremum}$$

Now fix n. for any $k \geq n$ we have $\epsilon_n \geq \epsilon_k$. By similiar consideration about the rectangles as above,

$$\int_E \tilde{f}_k\,dm \;\leq\; \mu_k(E^{\epsilon_n})$$

Let k to infinity, by MCT and the bound $\int |f_k - \tilde{f}_k| \leq \frac{1}{M_k}$ we get,

$$\mu(E) = \int_E f\,dm \;\leq\; \liminf_{k\to\infty} \mu_k(E^{\epsilon_n}) \tag{Infimum}$$

This by itself doesn't mean that

$$\mu(E) \;\leq\; \liminf_{k\to\infty} \mu_k(E)$$

However there is a remedey. Consider $\{\mu_k\}_{k=1}^{\infty}$. This set is tight, so by Prokhorov's theorem it has limit points. To prove $\mu_k \to \mu$ it suffices to prove that the only limit point of it is μ. Let $\mu_n \to u$. Then by Infimum and Supremum,

$$u(E) \leq \mu(E) \leq u(E^{\epsilon_n}) \qquad \forall n \in \mathbf{N} \quad \& \quad \forall\ E \in \mathcal{B}_{\mathbb{R}^n}$$

which means $u(E) = \mu(E)$ and hence $u = \mu$.

Proving $\nu_k \to \nu$ is exactly similiar and we are finally done with proposition3.