Introduction to Optimal Transport

Lectured by Matthew Thorpe

Lent 2019

Introduction to Optimal Transport
MWF / 12pm MR14 / and one extra lecture on 15th March / lecture on Monday 21st January is cancelled.
4 example classes(TBA), 4 example sheets.
Details and notes are available at www.damtp.ac.uk/people/mt748

Idea of the course :
1. Fundamental theorems, 2. Sepcial cases, 3. Wasserstien spaces, 4. Gradient flows, 5. Numerical methods.

References :
[1] Villani, 2003, "Topics in Optimal Transportation", would be the main reference
[2] Santambrogio, 2015, "Optimal Transport for Applied Mathematicians", would be the 2nd reference.
[3] Villani, 2008, "Optimal Transport : Old and New"
[4] Ambrosio, Gigli, Savaré, 2008, "Gradient Flows in Metric Spaces".
[5] Peyré, Cuturi, 2018, "Computational Optimal Transport"
[6] Doneri, Savaré, 2010, "Lecture Notes on Gradient Flows"
[7] S. Kolouri, S. Park, M. Thorpe, D. Slepčev and G. K. Rohde, "Optimal Mass Transport: Signal Processing and Machine-Learning Applications"

Chapters : 1. Background on Measure theory, 2. Foundation of Optimal Transport(Monge-Kantorovich formulation, existence of solution of the formulation), 3. Special Cases(e.g. 1D, discrete systems), 4. Kantorovich Duality(Optimal transport as a linear program, existence of solution of the dual problem, Kantorovich-Rubinstein Theorem), 5. Semi-discrete optimal transport, 6. Existence and characterisation of trasport maps(Knott-Smith optimality, Bremier's Theorem), 7. Wasserstein distances, 8. Gradient Flow in Wasserstein Spaces, 9. Numerical Approaches (although no programming is involved)

(18th January, Friday)

1 Background on Measure Theory

Definition) **σ-algebra** Σ on a space X is a collection of subsets of X such that (1) $X \in \Sigma$, (2) closed under complements, (3) closed under countable unions.

A **measure** μ is a function from Σ from Σ to $\mathbb{R} \cup \{+\infty\}$ satisfying (1) non-negativity, (2) null-empty sets, (3) countable additivity for disjoint sets.

A **probability measure** μ is a measure such that $\mu(X) = 1$. We denote the set of probability measures on X by $\mathcal{P}(X)$.

The **Borel σ-algebra** on topological space X is the smallest σ-algebra that contains all open sets in X. We denote the borel measures by $\mu : \mathcal{B}(X) \to [0, +\infty]$. (In this course, we will always assume that a measure is a Borel measure).

Space X will be **Polish space** if it is a complete, separable metric space.

Definition 1.1.) Let X be a metric space, and $C_b^0(X)$ be the continuous bounded function on X. Let $\{\mu_n\} \subset \mathcal{P}(X)$ and $\mu \in \mathcal{P}(X)$. We say $\mu \overset{*}{\to} \mu$ (**weak-* converges**) if

$$\int_X f(x) d\mu_n(x) \xrightarrow{n\to\infty} \int_X f(x) d\mu(x) \quad \forall f \in C_b^0(X)$$

Theorem 1.2.) *(Portmanteau theorem)* Let X be a metric space, $\{\mu_n\}_{n=1}^{\infty} \subset \mathcal{P}(X)$, $\mu \in \mathcal{P}(X)$. Then TFAE :

- $\lim_{n\to\infty} \int_X f(x) d\mu_n = \int_X f(x) d\mu \quad \forall f \in C_b^0(X)$.
- $\lim_{n\to\infty} \int_X f(x) d\mu_n = \int_X f(x) d\mu$ for all f Lipschitz and bounded.
- $\limsup \int f(x) d\mu_n \leq \int_X f(x) d\mu(x)$ for all f upper semi-continuous and bounded above.
- $\liminf \int f(x) d\mu \geq \int f(x) d\mu$ for all f lower semi-continuous and bounded below.
- $\limsup \mu_n(C) \leq \mu(C)$ for all closed sets C.
- $\liminf \mu_n(O) \geq \mu(O)$ for all open sets O.
- $\lim \mu_n(A) \geq \mu(A)$ for all continuity sets A of μ, *i.e.* $\mu(\partial A) = 0$.

Definition) $\mathcal{K} \subset \mathcal{P}(X)$ is **tight** if $\forall \epsilon > 0$, $\exists K_\epsilon \subset X$ compact such that $\mu(X \backslash K_\epsilon) < \epsilon$ for all $\mu \in \mathcal{K}$.

Note, if $\mathcal{K} = \{\mu\}$ and μ is *inner regular*, then $\mathcal{K}$ is tight.

Theorem 1.3) *(Prokhorov's Theorem)* Let X be a *Polish space*. Then a set $\mathcal{K} \subset \mathcal{P}(X)$ is tight *iff* the closure of $\mathcal{K}$ is sequentially compact in weak* topology.

Definition 1.4) Let $\mu \in \mathcal{P}(X)$ and $T : X \to Y$ be measurable. Then the **pushforward** of μ by T, $T_\#\mu = \nu$, is defined by

$$\nu(B) = \mu(T^{-1}(B)) \quad \forall B \text{ measurable}$$

If T is invertible, then $\nu(T(A)) = mu(A)$ for all A measurable.

Proposition 1.5) Let $\mu \in \mathcal{P}(X)$, $T : X \to Y$, $S : Y \to Z$ and $f \in L^1(Y)$. Then

1. *(Change of variables)* $\int_Y f(y) d(T_\#\mu)(y) = \int_X f(T(x)) d\mu(x) \quad \cdots\cdots\cdots \ (2.2)$
2. *(Composition of measures)* $(S \circ T)_\#\mu = S_\#(T_\#\mu)$.

proof)

1. Let $f \geq 0$. Then

$$\int_Y f(y) d(T_\#\mu)(y) = \sup\left\{ \int_Y s(y) d(T_\#(\mu))(y) \ : \ 0 \leq s \leq f, s \text{ simple}\right\}$$

If $s(y) = \sum_{i=1}^N a_i \delta_{U_i}(y)$ where $a_i = s(y)$ for all $y \in U_i$, we have

$$\int_Y s(y) d(T_\# \mu)(y) = \sum_{i=1}^N a_i T_\# \mu(y)$$

Let $V_i = T^{-1}(U_i)$ and $r = \sum_{i=1}^N a_i \delta_{V_i}$ then

$$\sum_{i=1}^N a_i \mu(V_i) = \int_X r(x) d\mu(x)$$

For $\forall x \in V_i$, we have $T(x) \in U_i$ and $r(x) = a_i = s(T(x)) \le f(T(x))$ and therefore

$$\sup_{0 \le s \le f} \int_Y s(y) d(T_\# \mu)(y) = \sup_{0 \le r \le f \circ T} \int_X r(x) d\mu(x)$$

and hence (2.2) holds for $f \ge 0$.

For general $f \in L^1(Y)$, decompose $f = f_+ - f_-$ then we have the result.

2. Let $A \subset Z$, then $T^{-1}(S^{-1}(A)) = (S \circ T)^{-1}(A)$ so

$$\begin{aligned} S_\#(T_\# \mu)(A) =& T_\# \mu(S^{-1}(A)) = \mu(T^{-1}(S^{-1}(A))) \\ =& \mu((S \circ T)^{-1}(A)) = (S \circ T)_\# \mu(A) \end{aligned}$$

(End of proof) □

Theorem 1.6) *(Disintegration of Measures)* Let $\mathbb{X}, Z$ be Polish spaces and $P : \mathbb{X} \to Z$ be measurable. Let $\pi \in \mathcal{P}(\mathbb{X})$ and define $\omega = P_\# \pi \in \mathcal{P}(Z)$. Then $\exists \omega$-almost everywhere uniquely defined family of probability measures $\{\pi(\cdot|z)\}_{z \in Z}$ such that $\pi(\cdot|z) \in \mathcal{P}(P^{-1}(z))$ and

$$\int_{\mathbb{X}} f(x) d\pi(x) = \int_Z \int_{P^{-1}(z)} f(x) d\pi(x|z) d\omega(z), \quad \forall f : \mathbb{X} \to [0, +\infty] \text{ measurable}$$

Comments :

(1) $\pi(A|z)$ can be thought of as the conditional probability of A given z.

(2) Usual application is when $\mathbb{X} = X \times Y$ and $\pi \in \Pi(\mu, \nu)$, $P(x, y) = x$, then $\mu = P_\# \pi$ and $P^{-1}(x) = \{x\} \times Y$ (with a abuse of notation, $\pi(\cdot|x) \in \mathcal{P}(Y)$) and we can write

$$\int_{X \times Y} f(x, y) d\pi(x, y) = \int_X \int_{\{x\} \times Y} f(x, y) d\pi(y|x) d\mu(x) \quad \forall f : X \times Y \to [0, +\infty] \text{ measurable}$$

(23rd January, Wednesday)

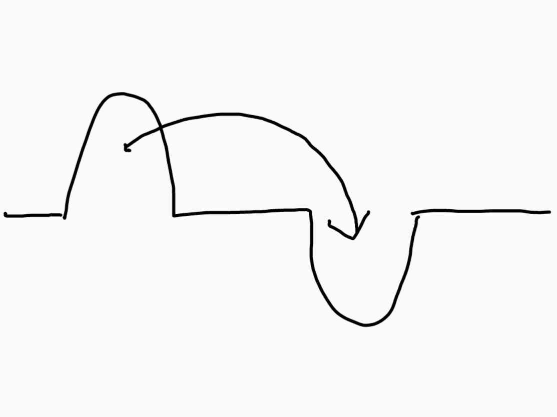
2 Formulation of Optimal Transport

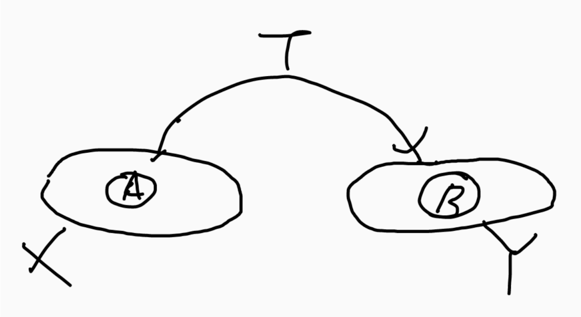
2.1 The Monge Formulation

Let μ, ν be measures on X, Y and $C : X \times Y \to [0, \infty]$ be the cost function.

Definition 2.1) We say that $T : X \to Y$ **transports** $\mu \in \mathcal{P}(X)$ to $\nu \in \mathcal{P}(Y)$(and we call T a **transport map**) if

$$\nu(B) = \mu(T^{-1}(B)) \quad \forall \nu - \text{measurable } B \quad \cdots\cdots\cdots \ (2.1)$$





If T is injective, then this is equivalent to saying that $\nu(T(A)) = \mu(A)$ for all μ-measurable sets A.

Let $\mu \in \mathcal{P}(X)$, $T : X \to Y$.

Definition 2.2) $\mu \in \mathcal{P}(X)$, $T : X \to Y$ be measurable. Then $T_\# \mu$ as defined in (2.1) is called the **pushforward** of μ by T.

Existence of Transport Map

Let μ, ν be probability measures as above. Then can we find $T : X \to Y$ such that $T_\# \mu = \nu$?

Examples :

- If $\mu = \delta_{x_1}$, $\nu = \frac{1}{2}(\delta_{y_1} + \delta_{y_2})$, we have $\frac{1}{2} = \mu(T^{-1}(y_1)) \in \{0, 1\}$, so we do not have such transport map.
- If in the discrete case with $|X| = |\{x_1, \cdots, x_n\}| = |Y| = |\{y_1, \cdots, y_n\}$, and $\mu = \frac{1}{n}\sum_i \delta_{x_i}$, $\nu = \frac{1}{n}\sum_i \delta_{y_i}$, then we always have the transport.(to be seen)
- If μ and ν are absolutely continuous w.r.t. the Lebesgue measures on X, Y, then the transport exists.(to be seen)

The Monge's Problem

Definition 2.4) *(Monge's Optimal Transport Problem)* Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ be given, then **Monge's optimal transport problem** asks to find

$$\text{minimize}_T \ \mathbb{M}(T) = \int_X C(y, T(x)) d\mu(x) \quad \text{subject to } \nu = T_\# \mu$$

If we assume that both measures are absolutely continuous w.r.t. the Lebesgue measure with $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$ and T is bijective with T, T^{-1} both differentiable, then (2.1) is equivalent to having $f(x) = g(T(y))|\det(\nabla T(x))|$. This is highly non-linear. So even in the simplest setting, it is often difficult to solve the Monge's problem.

2.2 The Kantorovich Formulation

Consider $\pi \in \mathcal{P}(X \times Y)$. We want $d\pi(x, y)$ to be the amount of mass transferred from x to y. We would have constraints $\pi(A, Y) = \mu(A)$ and $\pi(X, B) = \nu(B)$ for all measurable sets $A \subset X$, $B \subset Y$, *i.e.* the amount of mass transferred to and from a specific point is fixed. Denote the set of such π by $\Pi(\mu, \nu)$
Note, $\Pi(\mu, \nu)$ can never be empty, as the product measures $\mu \otimes \nu$ is in $\Pi(\mu, \nu)$.

Definition) *(Kantorovich Optimal Transport Problem)* Given $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$, the **Kantorovich optimal transport problem** asks to find

$$\text{minimize}_{\pi \in \Pi(\mu,\nu)} \ \mathbb{K}(\pi) = \int_{X \times Y} C(x, y) d\pi(x, y)$$

Obviously, the Monge's problem and the Kantorovich's problems are not equivalent - the existence of transport map T is not guaranteed for the Monge's problem in the first place. However, we can establish an inequality relation between $\inf \mathbb{K}\pi$ and $\inf \mathbb{M}(T)$.

Let us assume that $\exists T^\dagger : X \to Y$ is optimal in the Monge's problem, and let $d\pi(x, y) = d\mu(x)\delta_{y=T^\dagger x}$. Then

$$\begin{aligned}
\pi(A \times Y) &= \int_A \delta_{T^\dagger(x) \in Y} d\mu(x) = \mu(A) \\
\pi(X \times B) &= \int_X \delta_{T^\dagger(x) \in B} d\mu(x) = \mu((T^\dagger)^{-1}(B)) = T^\dagger_{\#}\mu(B) = \nu(B)
\end{aligned}$$

and

$$\int_{X \times Y} C(x, y) d\pi(x, y) = \int_X \int_Y C(x, y)\delta_{y=T^\dagger x} dy d\mu(x) = \int_X C(x, T^{-1}x) d\mu(x)$$

so we have $\inf \mathbb{K}(\pi) \leq \inf \mathbb{M}(T)$.

(25th January, Friday)

Last time, we have seen two formulations of optimal transport problem :

$$\begin{aligned}
&\min_{T : T_{\#}\mu = \nu} \int_X c(x, T(x)) d\mu(x) \quad \text{(Monge) ;} \\
&\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y) d\pi(x, y) \quad \text{(Kantorovich)}
\end{aligned}$$

If we find an optimal plan $\pi^\dagger : d\pi^\dagger(x, y) = d\mu(x)\delta_{y=T(x)}$, then T is optimal transport map and $\inf \mathbb{M}(T) \leq \inf \mathbb{K}(\pi)$. So combined with the inequality seen earlier, under a sufficient setting, the inequality is in fact an equality.

Interpolations

Monge : $\mu_t := \Big((1-t)id + tT^\dagger\Big)_\# \mu$. (compare with $\mu_t = (1-t)\mu + tv$)
Kantorovich : let $\mu = \sum_{i=1}^m \alpha_i \delta_{x_i}$, $\nu = \sum_{j=1}^n \beta_i \delta_{u_j}$, $c_{ij} = c(x_i, y_j)$ then the problem is written as

$$\min_\pi \sum_{i=1}^m \sum_{j=1}^n c_{ij}\pi_{ij}$$

$$\text{subject to} \quad \pi_{ij} \geq 0,\ \sum_{i=1}^m \pi_{ij} = \beta_j,\ \sum_{j=1}^n \pi_{ij} = \alpha_i$$

This linear program. The advantage for this setting is that it can also be written as

$$\inf_{\pi \geq 0, c^T\pi = (\mu^T, \nu^T)^T} c \cdot \pi = \sup_{C(\varphi^T, \phi^T) \leq c} (\mu \cdot \varphi + \nu \cdot \phi)$$

[Write in more detail later]

2.3 Existence of Transport Plans

Proposition 2.6) Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, X, Y Polish spaces and $c : X \times Y \to [0, \infty]$ is lower semi-continuous. Then $\exists \pi^\dagger \in \Pi(\mu, \nu)$ such that

$$\mathbb{K}(\pi^\dagger) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi)$$

proof) $\Pi(\mu, \nu)$ is non-empty. Let $\delta > 0$ and pick $K \subset X$, $L \subset Y$ compact so that

$$\mu(X \backslash K) \leq \delta, \quad \nu(Y \backslash L) \leq \delta$$

Take $(x, y) \in (X, Y) \backslash (K \times L)$. Then either $x \notin K$ or $y \notin L$, hence $(x, y) \in X \times (Y \backslash L)$ or $(x, y) \in (X \backslash K) \times Y$. Therefore, $\forall \pi \in \Pi(\mu, \nu)$, we have

$$\begin{aligned} \pi((X \times Y) \backslash (K \times L)) &\leq \pi(X \times (Y \backslash L)) + \pi((X \backslash K) \times Y) \\ &= \nu(Y \backslash L) + \mu(X \backslash K) \leq 2\delta \end{aligned}$$

so we see that $\Pi(\mu, \nu)$ is tight.

We want to see that $\Pi(\mu, \nu)$ is weak-* closed. So consider any weakly-* converging seqeunce $\pi_u \in \Pi(\mu, \nu) \xrightarrow{w-*} \pi \in M(X \times Y)$, *i.e.* $\forall f \in C_b^0(X \times Y)$, $\int_{X \times Y} f(x, y) d\pi_u(x, y) \to \int_{X \times Y} f(x, y) d\pi(x, y)$. Let $f(x, y) = \tilde{f}(x)$ be continuous and bounded. Then

$$\int_{X \times Y} f(x, y) d\pi_u(x, y) = \int_X \tilde{f}(x) d\mu(x) \to \int_X f(x) d\pi(x, y) = \int_X \tilde{f}(x) dP_\#^X \pi(x)$$

where $P^X(x, y) = x$ is the projection. Hence for all $\tilde{f} \in C_b^0(X)$, $\int_X \tilde{f}(x) d\mu(x) = \int_X \tilde{f}(x) dP_\#^X \pi(y)$ so $P_\#^X = \mu$. Similarly, $P_\#^Y \pi = \nu$ so we see that $\pi \in \Pi(\mu, \nu)$. Therefore, Π is weakly-* closed subset of $\mathcal{M}(X, Y)$.

Let $\pi_n \in \Pi(\mu, \nu)$ be a minimising sequence, *i.e.* $\mathbb{K}(\pi_n) \to \inf_{\pi \in \Pi} \mathbb{K}(\pi)$. Then by *Prohorov's theorem* and weak-* closedness of $\Pi(\mu, \nu)$, we may find $\pi^\dagger \in \Pi(\mu, \nu)$ and a subsequence π_{n_j} with $\pi_{n_j} \to \pi^\dagger$. Since c is bounded from below and continuous, we have

$$\int_{X \times Y} c(x, y) d\pi^\dagger(x, y) \leq \liminf_{j \to \infty} \int_{X, Y} c(x, y) d\pi_{n_j}(x, y) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi)$$

(End of proof) □

3 Special cases

We now just consider the special cases in which :

1. μ, ν on the real line *or*,

2. μ, ν on the discrete space

We will generalise the conditions on μ and ν later on based on the observations obtained in this chapter.

3.1 Optimal Transport in 1D

Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with culmulative distribution functions(CDF) F and G respectively. Then F, G are right-continuous, non-decreasing, $F(-\infty) = 0$ and $F(+\infty) = 1$. Define

$$\begin{aligned} & F^{-1}(t) := \inf\{x \in \mathbb{R} : F(x) > t\} \\ \Rightarrow \quad & F^{-1}(F(x)) \geq x, \quad F(F^{-1}(t)) \geq t \end{aligned}$$

If F is invertible, these inequalities are in fact equalities.

Theorem 3.1) Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with culmulative distribution functions F and G resp. Let $c(x, y) = d(x - y)$, for a function d convex and continuous. Let $\pi^\dagger \in \mathcal{P}(\mathbb{R}^2)$ with culmulative distribution function $H(x, y) = \min\{F(x), G(y)\}$. Then $\pi^\dagger \in \Pi(\mu, \nu)$, $\pi^\dagger$ optimal for the Kantorovich problem with cost function $c(x, y)$. Moreover,

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \int_0^1 d(F^{-1}(t) - G^{-1}(t)) dt$$

Corollary 3.2) Let μ and ν be as in the theorem.

(1) If $c(x, y) = |x - y|$, the optimal transport distance, we have

$$\inf_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \int_{\mathbb{R}} |F(x) - G(x)| dx$$

2) If μ does not give mass to atoms, then

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \min_{T : T_\# \mu = \nu} \mathbb{M}(T)$$

and $T^\dagger = G^{-1} \circ F$ is a minimiser to Monge's optimal transport prolem, *i.e.* $T^\dagger_\# \mu = \nu$ and $\inf_{T:T_\#\mu=\nu} \mathbb{M}(T) = \mathbb{M}(T^\dagger)$.

(28th January, Monday)

(Example Classes : 7th February, 21st February, 7th March and one next term, Thursdays)

Recap :

Definition 2.2) *(Monge Optimal Transport Problem (MOT))* Given $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, minimize $\mathbb{M}(T) = \int_X c(x, T(x))d\mu(x)$ over all measurable $T : X \to Y$ such that $T_{\#}\mu = \nu$.

Definition 2.3) *(Kantorovich Optimal Transport Problem (KOT))* Given $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, minimize $\mathbb{K}(\pi) = \int_{X \times Y} c(x, y)d\pi(x, y)$ over all $\pi \in \Pi(\pi, \nu)$.

If $T^{\dagger}$ minimizes MOT problem, then define $\pi^{\dagger} = (id \times T^{\dagger})_{\#}\mu \in \Pi(\mu, \nu)$, and so

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) \leq \mathbb{K}(\pi^{\dagger}) = \mathbb{M}(T^{\dagger})$$

and so $\min_{\pi} \mathbb{K}(\pi) \leq \inf_T \mathbb{M}(T)$.

This argument also works without assuming the existence of minimiser $T^{\dagger}$ by taking a sequence of T's such that $\mathbb{M}(T)$ converges to the minimum.

Proposition 2.4) Let $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$, X, Y Polish spaces, $c : X \times Y \to [0, \infty)$ lower semicontinuous(lsc), then $\exists \pi^{\dagger} \in \Pi(\mu, \nu)$ that minimises KOT.

> **proof)** Uses direct method from the calculus of variation. That is, the "compactness of lsc $\Rightarrow \exists$ minimiser".

Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$, with culmulative distribution function F, G respectively. We also defined the **generalized inverse** of a cdf F on $[0, 1]$ by

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

then $F^{-1}(F(x)) \geq x$ and $F(F^{-1}(t)) \geq t$.

Theorem 3.1) Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with cdf F, G respectively. Assume $c(x, y) = d(x, y)$ for some d convex (and continuous). Let $\pi^{\dagger}$ be the measure on $\mathbb{R}^2$ with cdf $H(x, y) = \min\{F(x), G(y)\}$. Then $\pi^{\dagger} \in \Pi(\pi, \nu)$ is a minimiser of KOT problem at the OT cost is

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \int_0^1 d(F^{-1}(t), G^{-1}(t))dt$$

> **Idea :** $H(x, y)$ has the following interpretation - we can not bring out from space X at x more than we have got $(F(x))$ and can not put in space Y at y than we need $(G(y))$.
>
> We will come back to the actual proof later.

Corollary 3.2) Under these assumptions of **Theorem 3.1**, the following hold :

1. If $c(x, y) = |x - y|$, then the KOT cost is the L^1 distance between cdf's, *i.e.*

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \int_{\mathbb{R}} |F(x) - G(x)|dx$$

2. If μ does not give mass to small sets (sets of Hausdorff dimension ¡ 1, or has no atoms, or its culmulative distribution function is continuous) then

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{F}(\pi) = \min_{T : T_{\#}\mu = \nu} \mathbb{M}(T)$$

and furthermore $T^{\dagger} = G^{-1} \circ F$ is a minimiser to the MOT problem.

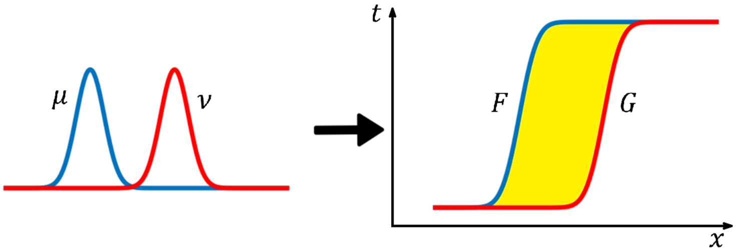
proof)

1. By **Theorem 3.1**, it is enough to show that

$$\int_0^1 |F^{-1}(t) - G^{-1}(t)|dt = \int_{\mathbb{R}} |F(x) - G(x)|dx$$

Define

$$A \subset \mathbb{R}^2 \text{ by } A = \{(x, y) : \min\{F(x), G((x)\} \leq t \leq \max\{F(x), G(x)\}, x \in \mathbb{R}$$



Note $A = \{(x, y) : \min\{F^{-1}(t), G^{-1}(t)\} \leq x \leq \max\{F^{-1}(t), G^{-1}(t)\}, t \in [0, 1]\}$ upto a set of Lebesgue measure 0. By Fubini's theorem,

$$\int_{\mathbb{R}} \int_{\min\{F(x),G(x)\}}^{\max\{F(x),G(x)\}} 1dtdx = \int_0^1 \int_{\min\{F^{-1}(t),G^{-1}(t)\}}^{\max\{F^{-1}(t),G^{-1}(t)\}} 1dxdt$$

so

$$\begin{aligned} &\int_{\mathbb{R}} \max\{F(x), G(x)\} - \min\{F(x), G(x)\}dx \\ &= \int_0^1 \max\{F^{-1}(t), G^{-1}(t)\} - \min\{F^{-1}(t), G^{-1}(t)\}dt \\ \Rightarrow \quad &\int_{\mathbb{R}} |F(x) - G(x)|dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt \end{aligned}$$

using $|a - b| = \max(a, b) - \min(a, b)$.

2. By composition of maps (**Proposition 2.3**) we have $T^\dagger_\# = G^{-1}_\#(F_\#\mu)$. We cliam (i) $G_\#\mathcal{L} = \nu$ and (ii) $F_\#\mu = \mathcal{L}$ where $\mathcal{L}$ is the Lebesgure measure on $[0, 1]$. Assumiing (i) and (ii), we have $T^\dagger_\#\mu = G^{-1}_\#\mathcal{L} = \nu$.

To prove (i),

$$\begin{aligned} G^{-1}_\#\mathcal{L}((-\infty, y]) &= \mathcal{L}(\{t : G^{-1}(t) \leq y\}) = \mathcal{L}(\{t : G(y) \geq t\}) \\ &= G(y) = \nu((-\infty, y]) \end{aligned}$$

and therefore $G^{-1}_\#\mathcal{L} = \nu$.

For (ii), since F is continuous (as μ does not give mass to atoms) so $\forall t \in (0, 1)$, $F^{-1}([0, t])$ is closed. Hence $\forall t \in (0, 1)$, $\exists x_t$ such that

$$F^{-1}([0, t]) = (-\infty, x_t] \quad \text{and} \quad F(x_t) = t$$

So $F_\#([0, t]) = \mu(x : F(x) \leq t) = \mu(x : x \leq x_t) = F(x_t) = t$. Hence $F_\#\mu = \mathcal{L}$.

Now by **Theorem 3.1**,

$$\begin{aligned}\inf \mathbb{K}(\pi) &= \int_0^1 d(F^{-1}(t) - G^{-1}(t))dt \\ &= \int_0^1 d(F^{-1}(t) - G^{-1}(t))d(F_{\#}\mu(t)) \\ &\geq \int_{\mathbb{R}} d(x - G^{-1}(F(x)))d\mu(x) \quad \text{(by } \textbf{Proposition 1.5}\text{)} \\ &= \int_{\mathbb{R}} d(x - T^{\dagger}(x))d\mu(x) \\ &= \mathbb{M}(T^{\dagger}) \geq \inf_{T:T_{\#}=0} \mathbb{M}(T) \geq \inf \mathbb{K}(\pi)\end{aligned}$$

where the last inequality follows from general relation between $\inf \mathbb{M}$ and $\inf \mathbb{K}$. This implies all inequalities are equalities, so $\min \mathbb{K}(\pi) = \min \mathbb{M}(T) = \mathbb{M}(T^{\dagger})$.

(End of proof) □

(30th January, Wednesday)

Example classes : MR11, 4pm - 5.5pm, (1) 7th Feb, (2) 21st Feb, (3) 7th March, one next term.

Recap : for $\mu, \nu \in \mathcal{P}(\mathbb{R})$, $F(x) = \int_{-\infty}^{x} d\mu(x)$, $G(y) = \int_{-\infty}^{y} d\nu(y)$.

We chage our notation, $F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}$... so we would have $F(F^{-1}(t)) \leq t$.

Theorem 3.1) $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with culmulative distribution function F, G respectively. Assume $c(x, y) = d(x - y)$ where d is convex and continuous. Let $\pi^{\dagger}$ be the measure on $\mathbb{R}^2$ with culmulative distribution function $H(x, y) = \min\{F(x), G(y)\}$. then $\pi^{\dagger} \in \Pi(\mu, \nu)$ and is a minimiser of KOT problem and the optimal cost is

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \int_0^1 d(F^{-1}(t) - G^{-1}(t))dt$$

The key idea in the proof of **Theorem 3.1** is monotonicity :

Definition) We say $\Gamma \subset \mathbb{R}^2$ **is monotone with respect to** d if $\forall (x_1, y_1), (x_2, y_2) \in \Gamma$, we have $d(x_1 - y_1) + d(x_2 - y_2) \leq d(x_1 - y_2) + d(x_2 - y_1)$.

This definition is given for 1D, but it can be generalised to higher dimensions. It is also used in convex analysis, *e.g.* the subdifferential of a convex function satisfies monotonicity property.

Intuition : the support of an optimal transport plan should be monotone. If $\text{supp}(\pi) = \Gamma$, then $(x_1, y_1), (x_2, y_2) \in \Gamma$ implies that mass is transported from x_i to y_i $(i = 1, 2)$. If $d(x_1 - y_1) + d(x_2 - y_2) \geq d(x_1 - y_2) + d(x_2 - y_2)$, *i.e.* cheaper to transport mass from x_1 to y_2 and x_2 to y_1, then to tranport from x_i to y_i $(i = 1, 2)$.

Example : Let $\Gamma = \{(x, y) : f(x) = y\}$ with f, d increasing. Then Γ is monotone. *(Exercise).*

Proposition 3.3) Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Assume $\pi^{\dagger} \in \Pi(\mu, \nu)$ is an optimal plan for the cost function $c(x, y) = d(x, y)$ where d is continuous. Then for all $(x_1, y_1), (x_2, y_2) \in \text{supp}(\pi)$, we have

$$d(x_1 - y_1) + d(x_2 - y_2) \leq d(x_1 - y_2) + d(x_2 - y_1)$$

i.e. supp(π) is *cyclically monotone.*

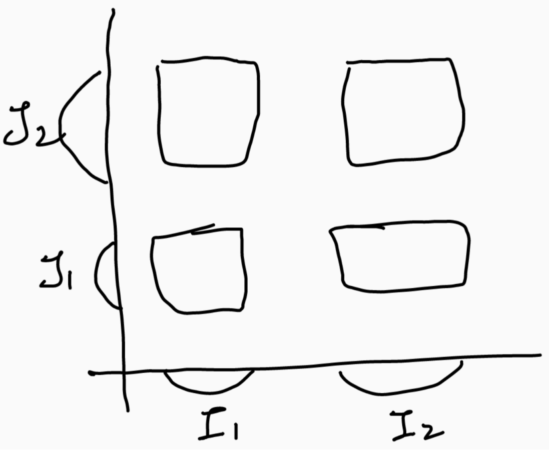
proof) Assume that $\exists \eta > 0$ such that

$$d(x_1 - y_1) + d(x_2 - y_2) - d(x_1 - y_2) - d(x_2 - y_1) > \eta$$

Let I_1, I_2, J_1, J_2 be closed intervals with the following properties :

$$\begin{cases} 1.\ x_i \in I_i, y_i \in J_i, \quad i = 1,2 \\ 2.\ |d(x-y) - d(x_i - y_j)| \le \epsilon \quad \text{for } x \in I_i, y \in J_i,\ i,j = 1,2, \text{ when } \epsilon < \eta/4 \\ 3.\ I_i \times J_j \text{ are disjoint} \\ 4.\ \pi^\dagger(I_1 \times J_1) = \pi^\dagger(I_2 \times J_2) = \delta > 0 \end{cases}$$

Properties 1-3 can be satisfied by chossing the intervals I_i, J_j small enough. Property 4 may not hold, but we at the moment assume it for simplicity. At the end of the proof, we will discuss how to avoid (4).



Idea : current plan $(\pi^\dagger)$ is $I_1 \mapsto J_1$, $I_2 \mapsto J_2$. We argue that $I_1 \mapsto J_2$, $I_2 \mapsto J_1$ provides a transport plan with lower cost.

Let $\tilde{\mu}_i = P^X_\# \pi^\dagger|_{I_i \times J_i}$, $\tilde{\nu}_i = P^Y_\# \pi^\dagger|_{I_i \times J_i}$, $i = 1,2$, where P^X is the projection to X and P^Y is the projection to Y. Choose any couplings $\tilde{\pi}_{12} \in \Pi(\tilde{\mu}_1, \tilde{\nu}_2)$, $\tilde{\pi}_{21} \in \Pi(\tilde{\mu}_2, \tilde{\nu}_1)$.

Define $\tilde{\pi}$ to satisfy

$$\tilde{\pi}(A \times B) = \begin{cases} \pi^\dagger(A \times B) & \text{if } (A \times B) \cap (I_i \times J_j) = \phi, \forall i,j \\ 0 & \text{if } A \times B \subset I_i \times J_i,\ i = 1,2 \\ \pi^\dagger(A \times B) + \tilde{\pi}_{12}(A \times B) & \text{if } A \times B \subset I_1 \times J_2 \\ \pi^\dagger(A \times B) + \tilde{\pi}_{21}(A \times B) & \text{if } A \times B \subset I_2 \times J_1 \end{cases}$$

(1st February, Friday)

proof continued) We check that $\tilde{\pi}$ defined above is in $\Pi(\mu, \nu)$

(i) for $B \subset \mathbb{R}$ with $B \cap (J_1 \cup J_2) = \phi$, we have $\tilde{\pi}(\mathbb{R} \times B) = \pi^\dagger(\mathbb{R} \times B) = \nu(B)$.

(ii) for $B \subset \mathbb{R}$ with $B \subset J_1$,

$$\begin{aligned}\tilde{\pi}(\mathbb{R} \times B) &= \tilde{\pi}((\mathbb{R} \backslash (I_1 \cup I_2)) \times B) + \tilde{\pi}(I_1 \times B) + \tilde{\pi}(I_2 \times B) \\ &= \pi^\dagger((\mathbb{R} \backslash (I_1 \cup I_2)) \times B) + 0 + \pi^\dagger(I_2 \times B) + \tilde{\pi}_{21}(I_2 \times B) \\ &= \pi^\dagger((\mathbb{R} \backslash I_1) \times B) + \pi^\dagger(I_1 \times B) \\ &= \pi^\dagger(\mathbb{R} \times B) = \nu(B)\end{aligned}$$

where the third equality follows because $\widetilde{\pi_{21}}(I_2 \times B) = \tilde{\nu}_1(B) = \pi^\dagger(I_1 \times B)$.

(iii) Similar computations when $B \subset J_2$ shows $\tilde{\pi}(\mathbb{R} \times B) = \nu(B)$ for all $B \subset \mathbb{R}$ measurable.
And analogously, $A \subset \mathbb{R}$, $\tilde{\pi}(A \times \mathbb{R}) = \mu(A)$. Hence $\tilde{\pi} \in \Pi(\mu, \nu)$.

Now since $\pi^\dagger$ and $\tilde{\pi}$ only differ on $(I_1 \cup I_2) \times (J_1 \cup J_2)$, we have

$$\begin{aligned}\mathbb{K}(\pi^\dagger) - \mathbb{K}(\tilde{\pi}) =& \int_{\mathbb{R}\times\mathbb{R}} d(x-y) d\pi^\dagger(x,y) - \int_{\mathbb{R}\times\mathbb{R}} d(x-y) d\tilde{\pi}(x,y) \\ =& \int_{(I_1\times J_1)\cup(I_2\times J_2)} d(x-y) d\pi^\dagger(x,y) \\ & - \int_{I_1\times J_2} d(x-y) d\tilde{\pi}(x,y) - \int_{I_2\times J_1} d(x-y) d\tilde{\pi}(x,y) \\ \geq& \delta(d(x_1-y_1)-\epsilon) + \delta((d(x_2-y_2)-\epsilon) - \delta(d(x_1-y_2)+\epsilon) - \delta(d(x_2-y_1)+\epsilon) \\ =& \delta\Big(d(x_1-y_1) + d(x_2-y_2) - d(x_1-y_2) - d(x_2-y_1) - 4\epsilon\Big) \\ \geq& \delta\Big(\eta - 4\epsilon\Big) > 0\end{aligned}$$

which contradicts with the fact that $\pi^\dagger$ is optimal. Hence, no such $\eta > 0$ exists.

Finally, as promised, let us discuss how we can modify the proof if Property 4 does not hold :

: Wlog, assume $\pi^\dagger(I_1 \times J_1) > \pi^\dagger(I_2 \times J_2)$. We let

$$\begin{aligned}\tilde{\mu}_1 &= \frac{\pi^\dagger(I_2 \times J_2)}{\pi^\dagger(I_1 \times J_1)} P^X_\# \pi^\dagger|_{I_1\times J_1} \\ \tilde{\nu}_1 &= \frac{\pi^\dagger(I_2 \times J_2)}{\pi^\dagger(I_1 \times J_1)} P^Y_\# \pi^\dagger|_{I_1\times J_1}\end{aligned}$$

and leave $\tilde{\mu}_2$, $\tilde{\nu}_2$ unchanged. Let $\tilde{\pi}$ be as before apart from when $A \times B \subset I_1 \times J_1$ - if $A \times B \subset I_1 \times J_1$, then we let

$$\tilde{\pi}(A \times B) = \pi^\dagger(A \times B)\Big(1 - \frac{\pi^\dagger(I_2 \times J_2)}{\pi^\dagger(I_1 \times J_1)}\Big)$$

Now the notations are more heavy, but nonetheless the theorem follows from exactly the same method.

(End of proof) □

Now we prove the main theorem of the chapter.

proof of Theorem 3.1) First, assume d is strictly convex. By **Proposition 2.4**, we may find $\pi^* \in \Pi(\mu, \nu)$ that is optimal for KOT problem, *i.e.* $\min_\pi \mathbb{K}(\pi) = \mathbb{K}(\pi^*)$. We show $\pi^* = \pi^\dagger$, where $\pi^\dagger$ is as in the statement of the theorem.

By **Proposition 3.3.**, $\Gamma = \text{supp}(\pi^*)$ is *monotone*. Let $(x_i, y_i) \in \Gamma$, $i = 1, 2$. We claim that if $x_1 < x_2$ then $y_1 \leq y_2$.

: Assume $y_1 > y_2$. Let $a = x_1 - y_1$, $b = x_2 - y_2$ and $\delta = x_2 - x_1 > 0$. We know from monotonicity

$$d(a) + d(b) \leq d(b - \delta) + d(a + \delta)$$

Let $t = \frac{\delta}{b-a} = \frac{x_2 - x_1}{(x_2 - x_1) + (y_1 - y_2)}$. Then $t \in (0, 1)$ *iff* $y_1 - y_2 > 0$. While

$$b - \delta = (1-t)b + tb - \delta = (1-t)b + t\Big(b - \frac{\delta}{t}\Big) = (1-t)b + t(b - b + a) = (1-t)b + ta$$

Similarly, $a + \delta = tb + (1-t)a$. So by strict convexity,

$$\begin{aligned} d(b - \delta) + d(a + \delta) <& (1-t)d(b) + td(a) + td(b) + (1-t)d(a) \\ =& d(b) + d(a) \end{aligned}$$

which contradicts monotonicity. Hence $t \notin (0, 1)$ so $y_1 \leq y_2$.

To show $\pi^\dagger = \pi^*$, we show $\pi^*((-\infty, x] \times (-\infty, y]) = \min\{F(x), G(y)\}$.

: Let $A = (-\infty, x] \times (y, \infty)$, $B = (x, \infty) \times (-\infty, y]$. Then $\pi^*(A)$ and $\pi^*(B)$ can not both be positive - indeed, if $\Gamma = \text{supp}(\pi^*)$, $(x_1, y_1) \in A \cap \Gamma$ and $(x_2, y_2) \in B \cap \Gamma$, then $x_1 < x_2$ and $y_2 > y_1$ which can not be the case by the result right above. This implies

$$\begin{aligned} \pi^*((-\infty, \pi] \times (-\infty, y]) &= \min\{\pi^*\Big(((-\infty, x] \times (-\infty, y]) \cup A\Big), \pi^*\Big(((-\infty, x] \times (-\infty, y])) \cup B\Big)\} \\ &= \min\{\pi^*\Big((-\infty, x] \times \mathbb{R}\Big), \pi^*\Big(\mathbb{R} \times (-\infty, y]\Big)\} \\ &= \min\{\mu((-\infty, x]), \nu(-\infty, y]\} \\ &= \min\{F(x), G(y)\} \end{aligned}$$

(4th February, Monday)

Theorem) Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ with culmulative distribution functions F and G respectively. Assume $c(x, y) = d(x - y)$ where $d : \mathbb{R} \to [0, +\infty)$ is convex (continuous). Let $\pi^\dagger$ be the measure on $\mathbb{R}^2$ with cdf $H(x, y) = \min\{F(x), G(y)\}$. Then $\pi^\dagger \in \Pi(\mu, \nu)$ and furthermore $\pi^\dagger$ is optimal for the KOT problem with cost c. Moreover, the OT cost is

$$\min_{\pi \in \Pi(\mu, \nu)} \mathbb{K}(\pi) = \int_0^1 d(F^{-1}(t) - G^{-1}(t))dt \quad \cdots\cdots\cdots \ (\star)$$

Last time : we proved the theorem assuming that d is strictly convex.

proof continued) Assume d is just convex. Since d is convex, it can be bounded below by an affine function say $d(x) \geq (ax + b)_+$ for some $a, b \in \mathbb{R}$. Then $f(x) = \frac{1}{2}\sqrt{4 + (ax + b)^2} + \frac{1}{2}(ax + b)$ is *strictly convex* and satisfies $0 \leq f(x) \leq 1 + d(x)$.

Define $d_\epsilon = d + \epsilon f$ so d_ϵ is strictly convex and satisfies $d \leq d_\epsilon \leq (1+\epsilon)d + \epsilon$. For $\pi \in \Pi(\mu, \nu)$ we have

$$\begin{aligned}\mathbb{K}(\pi^\dagger) &= \int_{\mathbb{R}\times\mathbb{R}} d(x-y)d\pi^\dagger(x,y) \leq \int_{\mathbb{R}\times\mathbb{R}} d_\epsilon(x-y)d\pi^\dagger(x,y) \\ &\leq \int_{\mathbb{R}\times\mathbb{R}} d_\epsilon(x-y)d\pi(x,y) \leq (1+\epsilon)\int_{\mathbb{R}\times\mathbb{R}} d(x-y)d\pi(x,y) + \epsilon \\ &= (1+\epsilon)\mathbb{K}(\pi) + \epsilon\end{aligned}$$

Let $\epsilon \to 0$, then has $\mathbb{K}(\pi^\dagger) \leq \mathbb{K}(\pi)$ for all $\pi \in \Pi(\mu,\nu)$, so $\pi^\dagger$ is optimal for KOT problem.

We now show $(\star)$ holds.

$\heartsuit$ **Claim** : $\pi^\dagger = (F^{-1}, G^{-1})_\# \mathcal{L}|_{[0,1]}$.

: To see this,

$$\begin{aligned}&(F^{-1}, G^{-1})_\# \mathcal{L}|_{[0,1]}((-\infty,x) \times (-\infty,y)) \\ =&\mathcal{L}|_{[0,1]}(F^{-1}, G^{-1})((-\infty,x) \times (-\infty,y)) \\ =&\mathcal{L}|_{[0,1]}(\{t : F^{-1}(t) > x, G^{-1}(t) > y\}) \\ =&\mathcal{L}|_{[0,1]}(\{t : F(x) > t, G(y) > t\}) \\ =&\min\{F(x), G(y)\} = \pi^\dagger((-\infty,x) \times (-\infty,y))\end{aligned}$$

Having the claim,

$$\begin{aligned}&\int_{\mathbb{R}\times\mathbb{R}} d(x-y)d\pi^\dagger(x,y) = \int_{\mathbb{R}\times\mathbb{R}} d(x-y)d(F^{-1}, G^{-1})_\# \mathcal{L}|_{[0,1]})(x,y) \\ =&\int_0^1 d(F^{-1}(t) - G^{-1}(t))dt \quad (x = F^{-1}(t), y = G^{-1}(t))\end{aligned}$$

by **Proposition 1.5**.

(End of proof) □

Remark 3.9 : We showed that if d is strictly convex and continuous, then the minimiser of the KOT problem is unique.

3.2 Existence of Transport Maps for Discrete Measures

Here we assume that $\mu = \frac{1}{n}\sum_{i=1}^m \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$.

Let $\sigma : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$ be any permutation. Then $T(x_i) = y_{\sigma(i)}$ is an admissible transport map.

Questions :

(1) Let $\nu = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ is it important that $m = n$? *Yes*, it is important. For example, let $\mu = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ and $\nu = \frac{1}{3}(\delta_{y_1} + \delta_{y_2} + \delta_{y_3})$, then there is no transport map between μ and ν.

(2) Do all $\{x_i\}, \{y_i\}$ need to be distinct? Yes for x_i's, and No for y_i's.

Definition) Let B be a convex and compact set in a Banach space M. The set of **extreme points**, denoted by $E(B)$, are the set of points in B that can not be written as non-trivial convex combinations of point in B, *i.e.* if $\pi \in B$, $\pi = \sum_{i=1}^m \alpha_i \pi_i$ where $\sum_{i=1}^m \alpha_i = 1$, $\alpha_i \in [0,1]$, and $\pi_i \in B$. If $\pi \in E(B)$, then $\alpha_i \in \{0,1\}$.

(6th February, Wednesday)

Recap : Let B be a convex and compact set in a Banach space m. Then $\pi \in B$ is an extremal point of B if $\pi = \sum_{i=1}^m \alpha_i \pi^{(i)}$, $\sum_{i=1}^m \alpha_i = 1$, $\alpha_i \geq 0$, $\pi^{(i)} \in B$ implies that $\alpha_i \in \{0,1\}$. We write $\pi \in E(B)$.

Theorem 3.5) *(Minkowski-Carathéodory Theorem)* Let $B \subset \mathbb{R}^m$ be non-empty, convex and compact. The for any $\pi^\dagger \in B$ there is a measures η supported on $E(B)$ such that for any affine function f,

$$f(\pi^\dagger) = \int f(\pi) d\eta(\pi)$$

proof) Proof is in the online notes.

Remark : The theorem can be generalised to Banach spaces where it is known as *Choquet's theorem.*

Theorem 3.6) *(Birkhoff's theorem)* Let B be the set of $n \times n$ bistochastic matrices, *i.e.*

$$B = \{\pi \in \mathbb{R}^{n\times n} : \pi_{ij} \geq 0, \sum_{j=1}^n \pi_i j = 1 \text{ for all } i, \sum_{i=1}^n \pi_{ij} = 1 \text{ for all } j\}$$

Then the set of extremal points of B is exactly the set of permutation matrices, i.e.

$$E(B) = \{\pi \in \{0,1\}^{n\times n} : \sum_{j=1}^n \pi_{ij} = 1 \text{ for all } i, \sum_{i=1}^n \pi_{ij} = 1 \text{ for all } j\}$$

proof) Proof is in the online notes.

We now prove the existence of optimal transport maps between $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_i}$.

Theorem 3.7) Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_i}$ and assume that $\{x_i\}_{i=1}^n$ are distinct ($\{y_i\}_{i=1}^n$ not necessarily distinct). Then there exists a solution to *Monge optimal transport problem* between μ and ν.

proof) Let $c_{ij} = c(x_i, y_j)$ and B be the set of $n \times n$ bistochastic matrices. Then KOT problem is

$$\text{minimize } \frac{1}{n}\sum_{i,j=1}^n c_{ij}\pi_{ij} \quad \text{over } \pi \in B$$

Note that we already know from **Proposition 2.4** that there is a minimiser (under extra assumptions) of KOT - here we do not use this result, but instead just use the idea of approximate minimisers. Let $\epsilon > 0$ and $\pi^\epsilon \in B$ such that

$$M \geq f(\pi^\epsilon) - \epsilon$$

where $M = \inf_{\pi \in B} f(\pi)$, and $f(\pi) = \sum_{i,j} c_{ij}\pi_{ij}$. Assume (for now, will come later) that B is convex and compact, then there exists a measure η supported on $E(B)$ such that

$$f(\pi^\epsilon) = \int f(\pi) d\eta(\pi)$$

We have

$$M \geq f(\pi^\epsilon) - \epsilon = \int_{E(B)} f(\pi) d\eta(\pi) - \epsilon \geq \inf_{\pi \in E(B)} f(\pi) - \epsilon \geq \inf_{\pi \in B} f(\pi) - \epsilon = M - \epsilon$$

So $\inf_{\pi \in E(B)} f(\pi) \in [M, M+\epsilon]$ for all $\epsilon > 0$. Let $\epsilon \to 0^+$, then we have $\inf_{\pi \in E(B)} f(\pi) = M$.

We assume (for now) that $E(B)$ is compact. Then using compactness, one may find a minimiser $\pi^\dagger \in E(B)$. Now by *Birkhoff's theorem*, we may find $\sigma \in \mathrm{Sym}(\{1, \cdots n\})$ such that $\pi^\dagger_{ij} = \delta_{j=\sigma(i)}$. Define $T^\dagger(x_i) = y_{\sigma(i)}$. Since $x_i \neq x_j$ for $i \neq j$, we have $T^\dagger$ is well-defined.

Let $T : \{x_i\}_{i=1}^n \to \{y_i\}_{i=1}^n$ be any transport map between μ and ν. Define $\pi_{ij} = \frac{1}{n}\delta_{j=\sigma(i)}$. Then we have $\pi \in B$. So

$$\begin{aligned}\mathbb{M}(T) = \int c(x, T(x)) d\mu(x) &= \frac{1}{n}\sum_{i=1}^n c(x_i, T(x_i)) = \sum_{i,j=1}^n c_{ij}\pi_{ij} \\ &\geq \sum_{i,j=1}^n c_{ij}\pi^\dagger_{ij} = \sum_{i=1}^n c(x_i, T^\dagger(x_i)) = \mathbb{M}(T^\dagger)\end{aligned}$$

so $T^\dagger$ minimises $\mathbb{M}$.

To finish the proof, we show (1) B is compact, (2) B is convex and (3) $E(B)$ is compact.

(1) Since all norms on finite dimensional space are equivalent, we choose here l^1-norm. Clearly B is bounded since for all $\pi \in B$, $\left\|\pi\right\|_1 = 1$. For closure, let $\pi^{(m)}$ converges to π. Then it follows $\pi^{(m)}_{ij} \to \pi_{ij}$ for each i, j, so $\pi_{ij} \geq 0$ and likewise $\sum_{j=1}^n \pi_{ij} = \lim_{n\to\infty} \sum_{j=1}^n \pi^{(m)}_{ij} = 1 = \sum_{i=1}^n \pi_{ij}$ so $\pi \in B$. Hence B is compact.

(2) Let $\pi^{(1)}$, $\pi^{(2)} \in B$, $t \in [0, 1]$ and let $\pi = t\pi^{(1)} + (1-t)\pi^{(2)}$ - then π is trivially in B.

(3) Since $E(B) \subset B$ hence bounded, it is enough to show that $E(B)$ is closed. Assume $\pi^{(m)} \to \pi$, where $(\pi^{(m)})_m \subset B$. Then $\pi^{(m)}_{i,j} \to \pi_{i,j}$, but $\pi^{(m)}_{ij} \in \{0, 1\}$ for each i, j, so $\pi_{i,j} \in \{0, 1\}$. So $\pi \in E(B)$.

(End of proof) □

(8th February, Friday)

4 Kantorovich Duality

Linear programmes have dual forms :

$$\text{(linear program)} \min_{y \geq 0, A^T y = c} b \cdot y = \sup_{Ax \leq b} c \cdot x \text{ (dual form)}$$

We saw that the KOT problem can be written as a linear program, if $\mu = \sum_{i=1}^m p_i \delta_{x_i}$, $\nu = \sum_{j=1}^n q_j \delta_{y_j}$ then

$$\min_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) = \min_{\pi \geq 0, A^T \pi = c} b \cdot \pi$$

where on the LHS, π is a $m \times n$ vector. On the RHS, b is the vector representing the cost of transport between x_i and y_j, A represents the projections onto the marginals and c represents the marginals.

So it would be unsurprising that the KOT problem has a dual form in the general set-up. We would see in this section :

4.1 Kantorovich Duality (statement of the main result)

4.2 Fenchel-Rockafellar Duality (min-max principle)

4.3 Proof of Kantorovich duality.

4.4 Existence of maximiser to the dual problem.

4.5 Kantorovich-Rubinstein theorem

4.1 Kantorovich Duality

Theorem 4.1) *(Kantorovich Duality, KD)* Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ when X, Y are Polish spaces. Let $c : X \times Y \to [0, +\infty]$ be lower semi-continuous. Define $\mathbb{J}$ by

$$\begin{aligned} \mathbb{J} : L^1(\mu) \times L^1(\nu) &\to \mathbb{R} \\ (\varphi, \psi) &\mapsto \int_X \varphi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \end{aligned}$$

and Φ_c by

$$\Phi_c = \{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu) : \varphi(x) + \psi(y) \leq c(x,y) \text{ for } \mu - \text{a.e. } x \in X, \nu - \text{a.e. } y \in Y\}$$

Then

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

The theorem is illustrated by the following short story.
"Shipper's Problem" : say we own factories and mines. We need to transport coal from our mines to our factories. Say the amount of coal produced at each min is fixed at the constant and at each factory is fixed, Say the cost of transporting coal from mine x to the factory y is $c(x, y)$. Then the optimal choice of transporting coal is

$$\min_{\pi \in \Pi(x,y)} \mathbb{K}(\pi)$$

where μ represents coal production on the mines at ν represents coal demand in factories.

Now enters the shipper who says they will transport coal for us and charge us for loading and unloading ships. In particular, we pay $\varphi(x)$ for loading at x and $\psi(y)$ for unloading at y. To make it in our interest the shippers says that $\varphi(x) + \psi(y) \leq c(x, y)$. If the shipper is clever, then he chooses (φ, ψ) such that we do not save anything.

sketch proof of Theorem 4.1) Let $M = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi)$ and

$$
\begin{aligned}
M = \inf_{\pi \in M_+(X\times Y)} \Big\{ \int_{X\times Y} c(x,y) d\pi(x,y) \\
+ \sup_{(\varphi,\psi)\in C_b^0(X)\times C_b^0(Y)} \Big(\int_X \varphi(x) d(\mu - P_\#^X \pi)(x) + \int_Y \psi(y) d(\nu - P_\#^Y \pi)(y) \Big) \Big\}
\end{aligned}
$$

This follows because

$$
\sup_{\varphi \in C_b^0(X)} \int_X \varphi(x) d(\mu - P_\#^X \pi)(x) = \begin{cases} 0 & \text{if } P_\#^X \pi = \mu \\ +\infty & \text{if otherwise} \end{cases}
$$

We can write

$$
\begin{aligned}
M = \inf_{\pi \in M_+(X\times Y)} \sup_{(\varphi,\psi)\in C_b^0(X)\times C_b^0(Y)} \Big\{ \int_{X\times Y} c(x,y) - \varphi(x) - \psi(y) d\pi(x,y) \\
+ \int_X \varphi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) \Big\}
\end{aligned}
$$

If we can justify min-max principle, we may exchange sup and inf so that,

$$
M = \sup_{(\varphi,\psi)\in C_b^0(X)\times C_b^0(Y)} \inf_{\pi \in M_+(X\times Y)} \Big\{ \int_{X\times Y} c(x,y) - \varphi(x) - \psi(y) d\pi(x,y) + \mathbb{J}(\varphi,\psi) \Big\}
$$

Assume that $\exists (x_0, y_0) \in X \times Y$ such that $c(x_0,y_0) - \varphi(x_0) - \psi(y_0) = -\epsilon < 0$ and let $\pi = \lambda \delta_{(x_0,y_0)}$, $\lambda > 0$. Then

$$
\begin{aligned}
&\inf_{\pi \in M_+(X\times Y)} \int_{X\times Y} c(x,y) - \varphi(x) - \psi(y) d\pi(x,y) \\
&\leq (c(x_0, y_0 - \varphi(x_0) - \psi(y_0))) \times \lambda \\
&= -\epsilon\lambda \to -\infty \quad \text{as } \lambda \to +\infty
\end{aligned}
$$

Hence we can restrict the supremum over $(\varphi,\psi) \in C_b^0(X) \times C_b^0(Y)$ to the subset for which $\varphi(x) + \psi(y) \leq c(x,y)$, hence $(\varphi,\psi) \in \Phi_c$. So for $(\varphi,\psi) \in \Phi_c$, has

$$
\inf_{\pi \in M_+(X\times Y)} \int_{X\times Y} c(x,y) - \varphi(x) - \psi(y) d\pi(x.y) \geq 0
$$

and by choosing $\pi \equiv 0$ we have

$$
\inf_{\pi \in M_+(X\times Y)} \int_{X\times Y} c(x,y) - \varphi(x) - \psi(y) d\pi(x,y) = 0
$$

So

$$
M = \sup_{(\varphi,\psi)\in\Phi_c} \mathbb{J}(\varphi,\psi)
$$

We still have to prove the min-max principle.

4.2 Fenchel Rockafellar Duality

The aim is to derive a rigorous min-max principle. We use methods from convex analysis. In particular, we use Legendre-Fenchel transform.

Definition 4.2) We define the **Legendre-Fenchel transformation**, as known as convex conjugate, of a convex function $\theta : E \to \mathbb{R} \cup \{+\infty\}$ where E is a normed vector space, is defined by

$$\begin{aligned} \Theta^* : E^* &\to \mathbb{R} \cup \{\infty\}, \\ z^* &\mapsto \sup_{z \in E}(\langle z^*, z\rangle - \Theta(z)) \end{aligned}$$

where E^* is the dual space of E.

The aim of this subsection is :

Theorem 4.3) *(Fenchel-Rockafellar duality)* Let E be a normed vector space and $\Theta, \Xi : E \to \cup\{+\infty\}$ two convex functions. Assume $\exists z_0 \in E$ such that $\Theta(z_0) < \infty$ and $\Xi(z_0) < \infty$ (and Θ is continuous at z_0). Then

$$\inf_{z\in E}(\Theta(z) + \Xi(z)) = \max_{z^* \in E^*}(-\Theta^*(-z^*) - \Xi^*(z^*))$$

In particular, the supremum on the RHS is attained.

(11th February, Monday)

Theorem 4.3) *(Fenchel-Rockafellar Dualtiy)* Let E be a normed vector space and $\Theta, \Xi : E \to \mathbb{R} \cup \{+\infty\}$ convex. Assume $\exists z_0 \in E$ such that $\Theta(z_0) < +\infty$, $\Xi(z_0) < +\infty$ and Θ is continuous at z_0. Then

$$\inf_{z\in E}(\Theta(z) + \Xi(z)) = \max_{z^* \in E^*}(-\Theta^*(-z^*) - \Xi^*(z^*))$$

Moreover, the supremum on the RHS is attained.

Some preliminary results :

Lemma 4.4) Let E be a normed vector space.

1. If $\Theta : E \to \mathbb{R} \cup \{+\infty\}$ is convex then so is the **epigraph** A defined by

   $$A = \{(z, t) \in E \times \mathbb{R} : t \geq \Theta(z)\}$$

2. If $\Theta : E \to \mathbb{R} \cup \{+\infty\}$ is concave then the **hypograph** B defined by

   $$B = \{(z, t) \in E \times \mathbb{R} : t \leq \Theta(z)\}$$

   is convex.

3. If $C \subset E$ is convex, then $\text{int}(C)$ is convex.

4. If $D \subset E$ is convex and $\text{int}(D) \neq \phi$ then $\overline{D} = \overline{\text{int}(D)}$

   proof) Exercise.

Theorem 4.5) *(Hahn-Banach)* Let E be a topological vector space. Assume A, B are convex non-empty and disjoint subsets of E and that A is open. Then there exists a closed hyperplane separating A and B.

proof) Not done here.

We now prove **Theorem 4.3**.

proof of Theorem 4.3) We may write

$$\begin{aligned}-\Theta^*(-z^*) - \Xi^*(z^*) &= \inf_{x,y\in E}\Big(\Theta(x) + \Xi(y) + \langle z^*, x-y\rangle\Big)\\ &\le \inf_{x\in E}\Big(\Theta(x)+\Xi(x)\Big)\end{aligned}$$

Taking supremum over $z^* \in E^*$ gives

$$\sup_{z^*\in E^*}\left(-\Theta^*(-z^*) - \Xi^*(z^*)\right) \le \inf_{x\in E}(\Theta(x)+\Xi(x))$$

To show the converse inequality, let $M = \inf_{x\in E}(\Theta(x)+\Xi(x))$. Define

$$A = \{(x,\lambda)\in E\times\mathbb{R} : \lambda \ge \Theta(x)\},\quad B = \{(y,\sigma)\in E\times\mathbb{R} : \sigma \le M - \Xi(y)\}$$

By **Lemma 4.4**, A and B are convex. By continuity of Θ at z_0, A has non-empty interior - so let $C = \mathrm{int}(A)$. By finiteness of Ξ at z_0, B is non-empty. If $(x,\lambda)\in C$, then $\lambda > \Theta(x)$ so $\lambda + \Xi(x) > \Theta(x) + \Xi(x) \ge M$, so $(x,\lambda)\notin B$. In particular, $B\cap C = \phi$. By the *Hahn-Banach theorem*, there is a hyperplanne $H = \{(x,\lambda) : \Phi(x,\lambda) = \alpha\}$ with Φ linear that separates B and C.

We may write $\Phi(x,\lambda) = f(x) + k\lambda$ where f is bounded linear and assume

$$\begin{cases}\forall(x,\lambda)\in C & f(x)+k\lambda > \alpha\\ \forall(x,\lambda)\in B & f(x)+k\lambda \le \alpha\end{cases}\qquad \cdots\cdots\cdots\ (\star)$$

(this follows from continuity of f, and $\overline{C} = A$) Since $(z_0,\lambda)\in A$ for λ large enough, has $k\ge 0$.

♣ **Claim** : $k > 0$

: Assume $k = 0$. Then $(\star)$ implies

$$\begin{cases}\forall(x,\lambda)\in A \Rightarrow f(x)\ge\alpha\\ \forall(x,\lambda)\in B\Rightarrow f(x)\le\alpha\end{cases} \Rightarrow \begin{cases}f(x)\ge\alpha & \forall x\in \mathrm{Dom}(\Theta)\\ f(x)\le \alpha & \forall x\in\mathrm{Dom}(\Xi)\end{cases}$$

(where $\mathrm{Dom}(\Theta) = \{x\in E : \Theta(x) < +\infty\}$, $\mathrm{Dom}(\Xi) = \{x \in E : \Xi(x) < +\infty\}$). Since $z_0 \in \mathrm{Dom}(\Theta)\cap\mathrm{Dom}(\Xi)$ then $f(z_0) = \alpha$. Since Θ is continuous at z_0, $\exists r > 0$ such that $B(z_0, r)\subset \mathrm{Dom}(\Theta)$. Hence $\forall |z|\le r$, $\delta\in(-1,1)$, we have (since $z_0 + \delta z \in \mathrm{Dom}(\Theta)$)

$$f(z_0+\delta z)\ge\alpha \xRightarrow{f \text{ linear}} f(z_0) + \delta f(z) \ge \alpha$$

and therefore $\delta f(z)\ge 0$. So $f(z) = 0$ for all $z\in B(0,r)$. As f was linear, we have $f\equiv 0$, $\Phi\equiv 0$. If $\alpha = 0$, then $H = E\times\mathbb{R}$ (recall, H was the separating hyperplane), and otherwise $H = \phi$ - in both cases, we get a contradiction.

Having the claim, we may choose $z^* = f/k$ as our candidate maximiser. By ($\star$),

$$\begin{aligned}\Theta^*(-f/k) =& \sup_{z\in E}(-\frac{f(z)}{k} - \Theta(z)) \\ =& -\frac{1}{k}\inf_{z\in E}(f(z) + k\Theta(z)) \leq -\alpha/k\end{aligned}$$

since $(z, \Theta(z)) \in A$. Similarly,

$$\begin{aligned}\Xi^*(f/k) =& \sup_{z\in E}(\frac{f(z)}{k} - \Xi(z)) \\ =& -M + \frac{1}{k}\sup_{z\in E}\Big(f(z) - k(M - \Xi(z))\Big) \leq -M + \alpha/k\end{aligned}$$

since $(z, M - \Xi(z)) \in B$. Hence,

$$\begin{aligned}M \geq& \sup_{z^*\in E^*}(-\Theta^*(-z^*) - \Xi^*(z^*)) \geq -\Theta^*(-f/k) - \Xi(f/k) \\ \geq& \frac{\alpha}{k} + M - \frac{\alpha}{k} = M\end{aligned}$$

so $M = \sup_{z^*\in E^*}(-\Theta^*(-z^*) - \Xi^*(z^*))$.

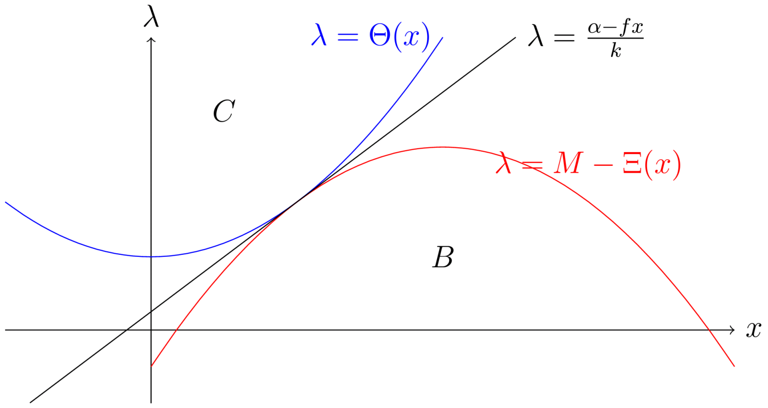
Note that the supremum is attained by $z^* = f/k$, so we have the 'moreover' part.

(End of proof) □

Remark 4.6. Why have we chosen $z^* = f/k$ as the maximiser? In terms of Lagrange multipliers, we can write

$$\inf_{x\in E}(\Theta(x) + \Xi(x)) = \inf_{x,y\in E}\sup_{z^*\in E^*}(\Theta(x) + \Xi(y) + \langle z^*, x - y\rangle)$$

Fix z^* then $\Theta(x) + \Xi(y) + \langle z^*, x - y\rangle$ is minimised when $\nabla\Theta(x) - z^*$ and $\nabla\Xi(y) = z^*$ Hence we want to find x, z^* such that $z^* = -\nabla\Theta(x) = \nabla\Xi(x)$.



(13th February, Wednesday)

4.3 Proof of Kantorovich Duality

We first prove that $\sup \mathbb{J} \leq \inf \mathbb{K}$ (easy part) and $\sup \mathbb{J} \geq \inf \mathbb{K}$ (hard part).

Lemma 4.7) Under the same conditions as **Theorem 4.1**, we have

$$\sup_{(\varphi,\psi)\in\Phi_c} \mathbb{J} \leq \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi)$$

proof) Let $(\varphi, \psi) \in \Phi_c$ and $\pi \in \Pi(\mu,\nu)$. Let $A \subset X$, $B \subset Y$ such that $\mu(A) = 1 = \nu(B)$ and $\varphi(x) + \psi(y) \leq c(x,y)$ for all $(x,y) \in A \times B$. Then $\pi(A \times B) = 1$(check) hence $\varphi(x) + \psi(y) \leq c(x,y)$ for π-a.e. (x,y). Then

$$\mathbb{J}(\varphi,\psi) = \int_{X\times Y} (\varphi(x) + \psi(y)) d\pi(x,y) \leq \int_{X\times Y} c(x,y) d\pi(x,y) = \mathbb{K}(\pi)$$

Taking supremum over $(\varphi,\psi) \in \Phi_c$ and infimum over $\pi \in \Pi(\mu,\nu)$ gives

$$\sup_{(\varphi,\psi)\in\Phi_c} \mathbb{J} \leq \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi)$$

(End of proof) □

Lemma 4.8) Under the same conditions as **Theorem 4.1**, has

$$\sup_{(\varphi,\psi)\in\Phi_c} \mathbb{J} \geq \inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi)$$

proof) We prove the lemma when X and Y are compact, and c is continuous. (if not, the proof get much harder)

Let $E = C_b^0(X\times Y)$ and equip E with the supremum norm. By *Riesz-Markov-Kakutani representation theorem*, has $E^* = \mathcal{M}(X \times Y)$ and the dual pairing

$$\langle \pi, u\rangle = \int_{X\times Y} u(x,y) d\pi(x,y) \quad \text{for } \pi \in E^*, u \in E$$

Define

$$\Theta(u) = \begin{cases} 0 & \text{if } u(x,y) \geq -c(x.y)\ \forall x, y \\ +\infty & \text{if otherwise} \end{cases}$$
$$\Xi(u) = \begin{cases} \mathbb{J}(\varphi,\psi) & \text{if } u(x,y) = \varphi(x) + \psi(y)\ \forall x, y \\ +\infty & \text{if otherwise} \end{cases}$$

(Note that Ξ is well-defined even though the representation $u = \varphi \oplus \psi$ is not unique.) Then it is easy to check that Θ and Ξ are convex.

For $u_0 \equiv 1$, then $u_0 \geq 0 \geq -c$, so $\Theta(u_0) = 0$ and Θ is continuous at u_0 and $\Xi(u_0) = 1 < \infty$. So by *Fenchel-Rockafellar Duality,*

$$\text{(A)} = \inf_{u\in E}(\Theta(u) + \Xi(u)) = \max_{\pi\in E^*}(-\Theta^*(-\pi) - \Xi^*(\pi)) = \text{(B)}$$

Here,

$$\begin{aligned} \text{(A)} \geq& \inf \left\{ \mathbb{J}(\varphi,\psi) : u \geq -c, u = \varphi\oplus\psi, \varphi \in L^1(\mu), \psi \in L^1(\nu) \right\} \\ =& \inf \left\{ \mathbb{J}(\varphi,\psi) : \varphi\oplus\psi \geq -c, \varphi \in L^1(\mu), \psi \in L^1(\nu) \right\} \\ =& -\sup_{\Phi_c} \mathbb{J}(\varphi,\psi) \end{aligned}$$

For (B),

$$\begin{aligned}\Theta^*(-\pi) =& \sup_{u\in E}\Big(-\int_{X\times Y} ud\pi - \Theta(u)\Big) = \sup_{u\geq -c, u\in E}\Big(-\int_{X\times Y} yd\pi\Big)\\ =& \sup_{u\leq c, u\in E}\int_{X\times Y} ud\pi = \begin{cases}\int_{X\times Y} cd\pi & \text{if } \pi\in M_+\\ +\infty & \text{if otherwise}\end{cases}\end{aligned}$$

and

$$\begin{aligned}\Xi^*(\pi) =& \sup_{u\in E}\Big(\int_{X\times Y} ud\pi - \Xi(u)\Big)\\ =& \sup_{u\in E, u=\varphi\oplus\psi}\Big(\int_{X\times Y} ud\pi - \int_X \varphi d\mu - \int_Y \psi d\nu\Big)\\ =& \sup_{u\in E, u=\varphi\oplus\psi}\Big(\int_X \varphi(x)d(P^X_\#\pi-\mu)(x) + \int_Y \psi(y)d(P^Y_\#\pi-\nu)(y)\Big)\\ =& \begin{cases}0 & \text{if } \pi\in\Pi(\mu,\nu)\\ +\infty & \text{if otherwise}\end{cases}\end{aligned}$$

So

$$(\text{B}) = \max_{\pi\in\Pi(\mu,\nu)} -\mathbb{K}(\pi) = -\inf_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)$$

So (A)=(B) implies the result.

(End of proof) □

4.4 Existence of Maximisers to the Dual Problem

The aim here is to find (φ,ψ) such that $\sup_{(\varphi,\psi)\in\Phi_c}\mathbb{J}(\varphi,\psi) = \mathbb{J}(\varphi^\dagger,\psi^\dagger)$.

Rather than working with the Legendre-Fenchel transform, it will be more convenient to use the closely related c-transform.

Definition 4.9) For $\varphi : X\to\overline{\mathbb{R}} := \mathbb{R}\cup\{+\infty\}$, $c: X\times Y\to\mathbb{R}$, the **c-transform**, φ^c and ψ^c are defined by

$$\begin{aligned}\varphi^c : Y\to\overline{\mathbb{R}}, &\quad y\mapsto \inf_{x\in X}(c(x,y)-\varphi(x))\\ \varphi^{cc} : X\to\mathbb{R}, &\quad x\mapsto \inf_{y\in Y}(c(x,y)-\varphi^c(y))\end{aligned}$$

If $Y = X^*$ and $c(x,y) = \langle y,x\rangle$ then $(-\varphi)^c(-y) = \varphi^c(y)$.

We can work similarly just using Legendre-Fenchel transformation but working just with Legendre-Fenchel transformation makes the problem complicated, and c-transform turns out to have different uses later on.

The main result :

Theorem 4.10) Let $\mu\in\mathcal{P}(X)$, $\nu\in\mathcal{P}(Y)$, X,Y are Polish spaces and $c: X\times Y\to[0,+\infty)$. Assume $\exists c_X\in L^1(\mu)$, $c_Y\in L^1(\nu)$ such that $c(x,y)\leq c_X(x)+c_Y(y)$ for μ-a.e. $x\in X$, ν-a.e. $y\in Y$. Then $\exists(\varphi^\dagger,\psi^\dagger)\in\Phi_c$ such that

$$\sup_{\Phi_c}\mathbb{J} = \mathbb{J}(\varphi^\dagger,\psi^\dagger)$$

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

The existence of c_X, c_Y in the statement of the theorem is effectively a condition on moments, *e.g.* $c(x,y) = |x-y|^2$, then $c(x,y) \le p|x|^p + p|y|^p$ so $c_X \in L^1(\mu)$ is equivalent to $\int |x|^p d\mu(x) < \infty$.

(15th February, Friday)

Recall : We defined **c-transform** by $\varphi^c(y) = \inf_{x\in X}(c(x,y) - \varphi(x))$. Our goal is to prove **Theorem 4.10.**

Lemma 4.11) Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, X, Y Polish spaces. For any $a \in \mathbb{R}$, $(\tilde{\varphi}, \tilde{\psi}) \in \Phi_c$, we have that $(\varphi, \psi) = (\tilde{\varphi}^{cc} - a, \tilde{\varphi}^c + a)$ satisfies $\mathbb{J}(\varphi,\psi) \ge \mathbb{J}(\tilde{\varphi}, \tilde{\psi})$ and $\varphi(x) + \psi(y) \le c(x,y)$ for μ-a.e. $x \in X$, ν-a.e. $y \in Y$.

Furthermore, if $\mathbb{J}(\tilde{\varphi}, \tilde{\psi}) > -\infty$ and there are $c_X \in L^1(\mu), c_Y \in L^1(\nu)$ such that $\varphi \le c_X$ and $\psi \le c_Y$ then $(\varphi,\psi) \in \Phi_c$.

proof) Clearly $\mathbb{J}(\varphi - a, \psi + a) = \mathbb{J}(\varphi, \psi)$ for all $a \in \mathbb{R}$ and for all $\varphi \in L^1(\mu)$, $\psi \in L^1(\nu)$. so it is enough to treat the case $a = 0$. In fact it is enough to show $\varphi = \tilde{\varphi}^{cc} \ge \tilde{\varphi}$, $\psi = \tilde{\varphi}^c \ge \tilde{\psi}$ and $\varphi(x) + \psi(y) \le c(x,y)$

1. $\psi(y) = \inf_{x\in X}(c(x,y) - \tilde{\varphi}(x)) \ge \tilde{\psi}(y)$ using that $\tilde{\varphi} + \tilde{\psi} \le c(x,y)$
2. $\varphi(x) = \inf_{y\in Y}(c(x,y) - \tilde{\varphi}^c(y)) = \inf_{y\in Y}\sup_{z\in X}(c(x,y) - c(z,y) + \tilde{\varphi}(z)) \ge \tilde{\varphi}(x)$ by choosing $z = x$.
3. $\varphi(x) + \psi(y) = \inf_{z\in Y}(c(x,z) - \tilde{\varphi}^c(z) + \tilde{\varphi}^c(y)) \le c(x,y)$ by choosing $z = y$..

For the furthermore part, we just need to show $\varphi \in L^1(\mu), \psi \in L^1(\nu)$. Let $M = \int_X c_X(x) d\mu(x) + \int_Y c_Y(y) d\nu(y) < +\infty$. Note

$$\int_Y (\varphi(x) - c_X(x)) d\mu(x) + \int (\psi(y) - c_Y(y)) d\nu(y) = \mathbb{J}(\varphi,\psi) - M \ge \mathbb{J}(\tilde{\varphi}, \tilde{\psi}) - M$$

By condition $\varphi \le c_X$ and $\psi \le c_Y$, $\int_Y (\varphi(x) - c_X(x)) d\mu = -\left\|\varphi - c_X\right\|_{L^1(\mu)}$ and $\int(\psi(y) - c_Y(y)) d\nu(y) = -\left\|\psi - c_Y\right\|_{L^1(\nu)}$ so this implies

$$\left\|\varphi - c_X\right\|_{L^1(\mu)} + \left\|\psi - c_Y\right\|_{L^1(\nu)} \le M - \mathbb{J}(\tilde{\varphi}, \tilde{\psi}) < +\infty$$

so

$$\begin{cases} \varphi - c_X \in L^1(\mu) \\ \psi - c_Y \in L^1(\nu) \end{cases} \quad \Rightarrow \quad \begin{cases} \varphi \in L^1(\mu) \\ \psi \in L^1(\nu) \end{cases}$$

as claimed.

(End of proof) □

Lemma 4.12) Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, X, Y Polish and $c : X \times Y \to \mathbb{R}$. Assume $c(x,y) \le c_X(x) + c_Y(y)$ for functions $c_X \in L^1(\mu)$, $c_Y \in L^1(\nu)$. Then there is a sequence $(\varphi_k, \psi_k)_k \subset \Phi_c$ such that $\mathbb{J}(\varphi_k, \psi_k) \to \sup_{\Phi_c} \mathbb{J}$ and satisfy $\varphi_k(x) \le c_X(x)$ for μ-a.e. x and $\psi_k(y) \le c_Y(y)$ for ν-a.e y for all $k \in \mathbb{N}$.

proof) Let $(\tilde{\varphi}_k, \tilde{\psi}_k) \in \Phi_c$ be a maximizing sequence. Note $\tilde{\varphi}_k$, $\tilde{\psi}_k$ are proper. Choose

$$a_k = \inf_{y \in Y}\{c_Y(y) - \tilde{\varphi}_k^c(y)\}$$

and define $(\varphi_k, \psi_k) = (\tilde{\varphi}_k^{cc} - a_k, \tilde{\varphi}_k^c + a_k)$. If $\varphi_k \leq c_X$ and $\psi_k \leq c_Y$ then by **Lemma 4.11**, we are done.

(1) **Claim :** $a_k < +\infty$

: Since $(\tilde{\varphi}_k, \tilde{\psi}_k) \in \Phi_c$, has $\tilde{\varphi}_k(x) \leq c(x,y) - \tilde{\psi}_k(y)$ for ν-a.e. $y \in Y$. Hence $\exists y_0 \in Y$ and $b_0 \in \mathbb{R}$ such that $\tilde{\varphi}_k(x) \leq c(x, y_0) + b_0$ so $\tilde{\varphi}_k^c(y_0) = \inf_{x \in X}(c(x,y_0) - \tilde{\varphi}_k(x)) \geq -b_0$. Hence $a_k \leq c_Y(y_0) - \varphi_c(y_0) \leq c_X(y_0) + b < +\infty$.

(2) **Claim :** $a_k > -\infty$.

: Has $c_Y(y) - \tilde{\varphi}_k^c(y) = \sup_{x \in X}(c_Y(y) - c(x,y) + \tilde{\varphi}_k(x)) \geq \sup_{x \in X}(-c_X(x) + \tilde{\varphi}_k(x)) \geq -c_X(x_0) + \tilde{\varphi}_k(x_0)$ for all x_0. So $a_k \geq \tilde{\varphi}_k(x_0) - c_X(x_0) > -\infty$.

(3) **Claim :** $\psi_k(y) \leq c_Y(y)$.

: Has $\psi_k(y) = \tilde{\varphi}_k^c(y) + a_k = \tilde{\varphi}_k^c(y) + \inf_{z \in Y}(c_Y(z) - \tilde{\varphi}_k^c(z)) \leq c_Y(y)$.

(4) **Claim :** $\varphi_k(x) \leq c_X(x)$.

: Has $\varphi_k(x) - c_X(x) = \inf_{y \in Y}(c(x,y) - \tilde{\varphi}_k^c(y)) - a_k - c_X(x) \leq \inf_{y \in Y}(c_Y(y) - \tilde{\varphi}_k^c(y)) - a_k = 0$.

(End of proof) □

We now prove the main theorem.

proof of Theorem 4.10) First note we have

$$\sup_{\varphi,\psi \in \Phi_c} \mathbb{J}(\varphi, \psi) \leq \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) \leq M := \int c_X d\mu + \int c_Y d\nu < +\infty$$

Let $(\varphi_k, \psi_k) \in \Phi_c$ be maximising sequence given by **Lemma 4.12**. Define $\varphi_k^{(l)}$ and $\varphi_k^{(l)}$ by

$$\begin{aligned}\varphi_k^{(l)}(x) &= \max\{\varphi_k(x) - c_X(x), -l\} + c_X(x)\\ \psi_k^{(l)}(y) &= \max\{\psi_k(y) - c_y(y), -l\} + c_Y(y)\end{aligned}$$

Note

$$\begin{cases} \varphi_k \leq \varphi_k^{(l)}, \quad \psi_k \leq \psi_k^{(l)}, \\ -l \leq \varphi_k^{(l)} - c_X(x) \leq 0, \quad -l \leq \psi_k^{(l)}(y) - c_Y(y) \leq 0 \\ \varphi_k^{(1)}(x) \geq \varphi_k^{(2)}(x) \geq \cdots, \quad \psi_k^{(1)}(x) \geq \psi_k^{(2)}(x) \geq \cdots \\ \varphi_k^{(l)}(x) + \psi_k^{(l)}(y) \leq \max\{c(x,y) - c_X(x) - c_Y(y), -l\} + c_X(x) + c_Y(y) \end{cases}$$

For each $l \in \mathbb{N}$, the sequence $\varphi_k^{(l)} - c_X(x)$ is bounded in L^∞. So $\varphi_k^{(l)} - c_X(x)$ is bounded in $L^p(\mu)$ for all $p \in (1, \infty)$. Since $L^p(\mu)$ is reflexive, we have "bounded + closed = weakly compact". Hence $\varphi_k^{(l)} - c_X$, upto a subsequence, converges weakly in $L^p(\mu)$ as $k \to \infty$. Since c_X is independent of k then $\varphi_k^{(l)}$, weakly in $L^p(\mu)$.

Choose $p = 2$. Then upto a subsequence, for each l, has $\varphi_k^{(l)} - c_X \xrightarrow{w} \varphi^{(l)} - c_X \in L^2(\mu) \subset L^1(\mu)$. By a diagonalization argument, we can assume that $\varphi_k^{(l)} - c_X \xrightarrow{w} \varphi^{(l)} - c_X$ for all $l \in \mathbb{N}$. Similarly we can choose a subsequence with $\psi_k^{(l)} - c_Y \xrightarrow{w} \psi^{(l)} - c_Y \in L^1(\nu)$ for each $l \in \mathbb{N}$ - and furthermore the limit is made pointwise a.e.

Since $c_X \geq \varphi^{(1)} \geq \varphi^{(2)} \geq \cdots$ and $\varphi^{(l)}$, $\psi^{(l)}$ are bounded above by L^1 function and monotonically decreasing, we can apply *monotone convergence theorem* to see

$$\lim_{l\to\infty} \int_X \varphi^{(l)}(x) d\mu(x) = \inf_X \varphi^+(x) d\mu(x), \quad \varphi^+(x) = \lim_{l\to\infty} \varphi^{(l)}(x)$$
$$\lim_{l\to\infty} \int_Y \psi^{(l)}(y) d\nu(x) = \inf_Y \psi^+(y) d\nu(y), \quad \psi^+(y) = \lim_{l\to\infty} \psi^{(l)}(y)$$

(needs modification)

(18th February, Monday)

Theorem 4.10) Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, X, Y Polish space, $c : X \times Y \to [0, \infty)$. Assume $\exists c_X \in L^1(\mu)$, $c_Y \in L^1(\nu)$ such that $c(x,y) \leq c_X(x) + c_Y(y)$ a.e. Then there is $(\varphi, \psi) \in \Phi_c$ such that $\sup_{\Phi_c} \mathbb{J} = \mathbb{J}(\varphi, \psi)$. Furthermore, we can choose $(\varphi, \psi) = (\eta^{cc}, \eta^c)$ for some $\eta \in L^1(\mu)$.

proof continued) From last time, we have (φ_k, ψ_k) maximising for $\mathbb{J}$. For each $l \in \mathbb{N}$ we found $\varphi_k \leq \varphi_k^{(l)}$, $\psi_k \leq \psi^{(l)}$ such that

$$\varphi_k^{(l)} + \psi_k^{(l)} \leq \max\{c(x,y) - c_X(x) - c_Y(y), -l\} + c_X(x) + c_Y(y) \quad \cdots\cdots\cdots \ (\oplus)$$
$$\varphi_k^{(l)} \xrightarrow{w} \varphi^{(l)}, \quad \psi_k^{(l)} \xrightarrow{w} \psi^{(l)} \quad \text{as } k \to \infty, \quad \text{in } L^1$$
$$c_x \geq \varphi^{(1)} \geq \varphi^{(2)} \cdots, c_Y \geq \psi^{(1)} \geq \psi^{(2)} \cdots$$

Now $\varphi^{(l)}$, $\psi^{(l)}$ are bounded above and monotonically decreasing, so by the monotone convergence theorem,

$$\lim_{l\to\infty} \int_X \varphi^{(l)}(x) d\mu(x) = \int_X \varphi^\dagger(x) d\mu(x), \quad \varphi^\dagger = \lim_{l\to\infty} \varphi^{(l)}(x)$$
$$\lim_{l\to\infty} \int_X \psi^{(l)}(x) d\nu(x) = \int_X \psi^\dagger(x) d\nu(x), \quad \psi^\dagger = \lim_{l\to\infty} \psi^{(l)}(y)$$

We now want to show : (1) $(\varphi^\dagger, \psi^\dagger) \in \Phi_c$ and (2) $\mathbb{J}(\varphi, \psi) \leq \mathbb{J}(\varphi^\dagger . \psi^\dagger)$ for all $(\varphi, \psi) \in \Phi_c)$.

For (1), letting $l \to \infty$ in $(\oplus)$, has $\varphi^\dagger(x) + \psi^\dagger(y) \leq c(x,y)$. To check integrability, since $\varphi^\dagger(x) \leq c_X(x)$, and $\psi^\dagger(y) \leq c_Y(y)$ so

$$\begin{aligned} & - \left\| \psi^\dagger - c_X \right\|_{L^1(\mu)} - \left\| \psi^\dagger - c_Y \right\|_{L^1(\nu)} \\ & = \int_X \varphi^\dagger(x) - c_X(x) d\mu(x) + \int_Y \psi^\dagger(y) - c_Y(y) d\nu(y) \\ & = \mathbb{J}(\varphi^\dagger, \psi^\dagger) - M \end{aligned}$$

where $M = \int_X c_X d\mu + \int_Y c_Y d\nu$. So $\left\| \varphi^\dagger - c_X \right\|_{L^1(\mu)} + \left\| \psi^\dagger - c_Y \right\|_{L^1(\nu)} \leq M - \mathbb{J}(\varphi^\dagger, \psi^\dagger) \leq M - \sup_{\Phi_c} \mathbb{J}(\varphi, \psi)$. So $\varphi^\dagger \in L^1(\mu)$ and $\psi^\dagger \in L^1(\nu)$. Hence $(\varphi^\dagger, \psi^\dagger) \in \Phi_c$.

For (2),

$$\begin{aligned}\sup_{\Phi_c}\mathbb{J}=\lim_{k\to\infty}\mathbb{J}(\varphi_k,\psi_k)&\le\lim_{k\to\infty}\mathbb{J}(\varphi_k^{(l)},\psi_k^{(l)})\quad\text{for any } l\\ &=\mathbb{J}(\varphi^{(l)},\psi^{(l)})\end{aligned}$$

Let $l\to\infty$, then has

$$\sup_{\Phi_c}\mathbb{J}\le\lim_{l\to\infty}\mathbb{J}(\varphi^{(l)},\psi^{(l)})=\mathbb{J}(\varphi^{\dagger},\psi^{\dagger})$$

For the furthermore statement, we use the c-transform trick as in **Lemma 4.12** : for all $a\in\mathbb{R}$, by **Lemma 4.12**,

$$\mathbb{J}(\varphi^{\dagger},\psi^{\dagger})\le\mathbb{J}((\varphi^{\dagger})^{cc}-a,(\varphi^{\dagger})^{c}+a)=\mathbb{J}((\varphi^{\dagger})^{cc},(\varphi^{\dagger})^{c})$$

Let $a=\inf_{y\in Y}(c_Y(y)-(\varphi^{\dagger})^c(y))$, one can show that $a\in\mathbb{R}$ as in the proof of **Leamma 4.12**, and $(\varphi^{\dagger})^c(y)+a\le c_Y(y)$ and $(\varphi^{\dagger})^{cc}(x)-a\le c_X(x)$. Hence,

$$((\varphi^{\dagger})^{cc}+a,(\varphi^{\dagger})^{c}-a)\in L^1(\mu)\times L^1(\nu)$$

so $((\varphi^{\dagger})^{cc},(\varphi^{\dagger})^{c})\in L^1(\mu)\times L^1(\nu)$.

(End of proof) □

4.5 Kantorovich-Rubinstein Theorem

Theorem 4.13) *(Kantorovich-Rubinstein Theorem)* Let $X=Y$ be Polish, $c:X\times Y\to[0,+\infty]$ be lower semi-continuous metric on X. Define $\left\|f\right\|_{Lip}=\sup_{x\neq y\in X}\frac{|f(x)-f(y)|}{c(x,y)}$ for $f:X\to\mathbb{R}$. Let $\mu\in\mathcal{P}(X)$, $\nu\in\mathcal{P}(X)$ and assume $c(\cdot,x_0)\in L^1(\mu)$, $c(x_0,\cdot)\in L^1(\nu)$ for some point $x_0\in X$. Then

$$\min_{\pi\in\Pi(\mu,\nu)}\mathbb{K}(\pi)=\sup\Big\{\int_X fd(\mu-\nu):f\in L^1(|\mu-\nu|),\ \left\|f\right\|_{Lip}\le 1\Big\}$$

proof) We only prove the theorem when c is bounded, *i.e.* $\sup_{x,y\in X}c(x,y)\le c<+\infty$. So for any f with $\left\|f\right\|_{Lip}\le 1$ we have $|f(x)-f(y)|\le c$. So f is bounded, and hence we always have $f\in L^1(|\mu-\nu|)$. By **Theorem 4.1**, it is enough to check

$$\sup\mathbb{J}=\sup\Big\{\int_X fd(\mu-\nu):\left\|f\right\|_{Lip}\le 1\Big\}$$

From now assume $\forall\eta\in L^1(\mu)$, we have (a) $\|\eta^c\|_{Lip}\le 1$, (b) $\eta^{cc}=-\eta^c$ and (c) for all f with $\|f\|_{Lip}\le 1$, $(-f,f)\in\Phi_c$.

Assuming (a), (b), (c), by **Theorem 4.10**,

$$\begin{aligned}\sup_{\Phi_c}\mathbb{J}&=\sup_{\eta\in L^1(\mu)}\mathbb{J}(\eta^{cc},\eta^c)\\ &=\sup_{\eta\in L^1(\mu)}\mathbb{J}(-\eta^c,\eta^c)\quad\text{by assumptoin (b)}\\ &\le\sup_{\|f\|_{Lip}\le 1}\mathbb{J}(-f,f)\quad\text{by assumption (a)}\\ &\le\sup_{(\varphi,\psi)\in\Phi_c}\mathbb{J}(\varphi,\psi)\end{aligned}$$

so $\sup_{\Phi_c} \mathbb{J} = \sup_{\|f\|_{Lip} \leq} \mathbb{J}(f, -f)$ and $\mathbb{J}(f, -f) = \int_X f d(\mu - \nu)$. So the theorem is proved assuming (a), (b), (c).

Now let us justify the assumptions.

(a) It is sufficient to show $|\eta^c(x) - \eta^c(y)| \leq c(x, y)$

$$\begin{aligned} \eta^c(x) - \eta^c(y) = & \sup_{z \in X} \inf_{w \in X} \Big(c(x, w) - c(y, z) + \eta(z) - \eta(w) \Big) \\ \leq & \sup_{z \in X} \Big(c(x, z) - c(y, z) \Big) \quad \text{by choosing } w = z \\ \leq & c(x, y) \quad \text{by triangle inequality} \end{aligned}$$

By symmetry, $|\eta^c(x) - \eta^c(y)| \leq c(x, y)$ for all x, y, so $\left\| \eta^c \right\|_{Lip} \leq 1$.

(b) Has $\eta^{cc}(x) = \inf_{y \in X}(c(x, y) - \eta^c(y)) \leq \eta^c(x)$ since $|\eta^c(y) - \eta^c(x)| \leq c(x, y)$ by (a). On the other hand, again by (a), $\eta^{cc}(x) \geq \inf_{y \in X}(c(x, y) - (\eta^c(x) + c(x, y))) = -\eta^c(x)$ so putting these together, we have $\eta^{cc} = -\eta^c$.

(c) For all f with $\|f\|_{Lip} \leq 1$, has $f \in L^\infty$ so $f \in L^1(\mu) \cap L^1(\nu)$ and by definition of being Lipschitz,

$$f(x) - f(y) \leq c(x, y)$$

so $(f, -f) \in \Phi_c$.

(End of proof) □

(20th February, Wednesday)

5 Semi-Discrete Optimal Transport

Assume

1. $\nu = \sum_{j=1}^n m_j \delta_{y_j}$, $\{m_i\}_{i=1}^n \subset [0, 1]$, $\sum_{j=1}^n m_j = 1$ and $\{y_j\}_{j=1}^n \subset \mathbb{R}^d$.
2. $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ has density ρ.

Definition 5.1) The *Laguerre diagram (power diagram)* for a set of points $\{y_j\}_{j=1}^n$ and weight $\{w_j\}_{j=1}^n \subset \mathbb{R}$ is the collection of sets

$$L_j = \{x \in \mathbb{R}^d : |x - y|^2 - w_j < |x - y_i|^2 - w_i \ \forall i \neq j\}$$

for $j = 1, \cdots, n$.

Comments :

- If $w_j = 0$, then the *Laguerre diagram* are Voronoi cells.
- Each L_i is open.

Aims :

- Show there exists optimal $T : \mathbb{R}^d \to \mathbb{R}^d$ that defines a Laguerre diagram.

- and the weights $\{w_j\}_{j=1}^n$ for the optimal Laguerre diagram solve a concave variational problem

$$\max g(w) \quad \text{where } w = (w_1, \cdots w_n)$$

where g is as defined in the next lemma.

Lemma 5.2) Let $\rho \in L^1(\mathbb{R}^d)$ be a probability density, $\{m_j\}_{j=1}^m \subset [0,1]$ satisfy $\sum_{j=1}^m m_j = 1$ and $\{y_j\}_{j=1}^m \subset \mathbb{R}^d$. Then $g : \mathbb{R}^n \to \mathbb{R}$ defined by

$$g(w) = \int_{\mathbb{R}^d} \inf_j (|x - y_j|^2 - w_j)\rho(x)dx + \sum_{j=1}^n w_j m_j \quad \cdots\cdots\cdots \ (\star)$$

is concave.

Idea of proof : Introduce $\gamma : \mathbb{R}^n \to \{1, \cdots, n\}$. Define

$$G(\gamma, w) = \int_{\mathbb{R}^d} \Big(|x - y_{\gamma(x)}|^2 - w_{\gamma(x)} \Big) \rho(x)dx + \sum_{j=1}^n w_j m_j$$

Note (1) $w \mapsto G(\gamma, w)$ is an affine function, so concave and (2) $g(w) = \inf_\gamma G(\gamma, w)$. So g is also concave.

Lemma 5.3) Define g by $(\star)$ for $\rho \in L^1(\mathbb{R}^d)$, $\{y_j\}_{j=1}^n \subset \mathbb{R}^n$, $\{m_j\}_{j=1}^n \subset \mathbb{R}$. Let $\{L_i(w)\}_{i=1}^n$ be a *Laugerre diagram* with weights w and points $\{y_j\}_{j=1}^n$. Then

$$\frac{\partial g}{\partial w_i}(w) = -\int_{L_i(w)} \rho(x)dx + m_i$$

sketch proof) Let $\alpha_j(x, w) = \chi_{L_j(w)}(x)(|x - y_j|^2 - w_j)\rho(x)$ so

$$g(w) = \sum_{j=1}^n \Big(\int_{\mathbb{R}^d} \alpha_j(x, w)dx + w_j m_j \Big)$$

For any $x \in L_i(w)$, we have $\chi_{L_i(w+\pm te_i)}(x) = \chi_{L_j(w)}(x)$ for $t > 0$ sufficiently small, where e_i is the unit vector in i-direction. Moreover,

$$\frac{1}{t}\Big(\alpha_j(x, w + te_i) - \alpha_j(x, w)\Big) = -\chi_{L_j(w)}(x)\delta_{ij}\rho(x)$$

Hence

$$\begin{aligned}
\frac{\partial g}{\partial w_i}(w) &= \lim_{t \to 0^+} \frac{1}{t}\Big(g(w + te_i) - g(w)\Big) \\
&= \sum_{j=1}^n \lim_{t \to 0^+} \left[\int_{\mathbb{R}^d} \frac{1}{t}\Big(\alpha_j(x, w + te_i) - \alpha_j(x, w)\Big)dx + (w_j + \delta_{ij}t)m_j - w_j m_j \right] \\
&= -\int_{\mathbb{R}^d} \chi_{L_i(w)}(x)\rho(x)dx + m_i
\end{aligned}$$

(End of proof) □

Then comes the main result of the section.

Theorem 5.4) Assume $\{y_j\}_{j=1}^m \subset \mathbb{R}$, $\{m_j\}_{j=1}^m \subset [0,1]$, $\sum_{j=1}^n m_j = 1$ and $\nu = \sum_{j=1}^n m_j \delta_{y_j}$. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ have density ρ and $g(w)$ is defined by $(\star)$, maximised by $w = (w_1, \cdots, w_n)$, and let $\{L_j\}_{j=1}^n$ be the corresponding *Laguerre diagram.* Now define

$$\begin{aligned} &T^\dagger(x) = y_j \quad \text{if } x \in L_j \quad \text{(which defines } \mu\text{-a.e.)} \\ &\psi^\dagger(y_j) = w_j, \quad \varphi^\dagger(x) = \inf_j (|x - y_j|^2 - w_j) \end{aligned}$$

Then,

1. $T^\dagger$ is a solution to the MOT problem with cost $c(x,y) = |x-y|^2$.
2. $(\varphi^\dagger, \psi^\dagger)$ are an optimal pair for the Kantorovich dual problem with cost $c(x,y) = |x-y|^2$.

proof) We first assume that $T^\dagger$, $\varphi^\dagger$ and $\psi^\dagger$ are admissible for the optimisation problem:

(a) $\varphi^\dagger \in L^1(\mu)$, (b) $T^\dagger_\# \mu = \nu$, (c) $\int_{L_j} \rho(x) dx = m_j$

Assume (a), (b), (c) then

$$\varphi^\dagger(x) + \psi^\dagger(y_i) = \inf_j \Big(|x - y_j|^2 - w_j \Big) + w_i \leq |x - y_i|^2 = c(x, y_i)$$

So $(\varphi^\dagger, \psi^\dagger) \in \Phi_c$. Now we have

$$\begin{aligned} \mathbb{M}(T^\dagger) &\geq \inf\nolimits_{T_\#\mu=\nu} \mathbb{M}(T) \quad \text{(by (b))} \\ &\geq \min\nolimits_{\pi \in \Pi(\mu,\nu)} \mathbb{K}(\pi) \\ &= \sup\nolimits_{(\varphi,\psi)\in\Phi)c} \mathbb{J}(\varphi, \psi) \quad \text{by Theorem 4.1} \\ &\geq (\varphi^\dagger, \psi^\dagger) \end{aligned} \qquad \cdots\cdots\cdots \ (\oplus)$$

and

$$\begin{aligned} \mathbb{J}(\varphi^\dagger, \psi^\dagger) &= \int_{\mathbb{R}^d} \varphi^\dagger(x)\rho(x)dx + \sum_{j=1}^n m_j \psi^\dagger(y) \\ &= \sum_{j=1}^n \Big(\int_{L_j} \varphi^\dagger(x)\rho(x)dx + m_j \psi^\dagger(y_j) \Big) \\ &= \sum_{j=1}^n \int_{L_j} \Big(\big(|x - y_j|^2 - w_j \big) \rho(x) dx + m_j w_j \Big) \\ &= \sum_{j=1}^n \int_{L_j} |x - y_j|^2 \rho(x) dx \\ &= \sum_{j=1}^n \int_{L-j} |x - T^\dagger(x)|^2 \rho(x) dx = \mathbb{M}(T^\dagger) \end{aligned}$$

Hence all inequalities in $(\oplus)$ are all equalities and in particular

$$\begin{aligned} \mathbb{M}(T^\dagger) &= \min_{T: T_\#\mu=\nu} \mathbb{M}(T) \\ \mathbb{J}(\varphi^\dagger, \psi^\dagger) &= \sup_{(\varphi,\psi)\in\Phi_c} \mathbb{J}(\varphi, \psi) \end{aligned}$$

Now we are left to prove assumptions (a), (b), (c).

(a) Has

$$-\sup_j w_j \leq \varphi^\dagger(x) \leq |x-y_i|^2 - w_i$$

for any i so $|\varphi^\dagger(x)| \leq 2|x|^2 + C$ for a constant $C = 2|y_1|^2 - w_1 + \sup_j w_j$ and

$$\left\|\varphi\right\|_{L^1(\mu)} \leq 2\int_{\mathbb{R}^d} |x|^2 d\mu(x) + C < +\infty$$

(b) Pick $i \in \{1, \cdots, n\}$. Then

$$\begin{aligned}\mu((T^\dagger)^{-1}(y_i)) =& \mu(\{x : T^\dagger(x) = y_i\}) = \mu(L_i) = m_i \quad \text{(by (c))} \\ =& \nu(\{y_i\})\end{aligned}$$

So $T^\dagger_\# \mu = \nu$ as required.

(c) Since w maximises g, has $\frac{\partial g}{\partial w_i}(w) = 0$ for each i, but $\frac{\partial g}{\partial w_i}(w) = -\int_{L_j} \rho(x)dx + m_j$ by **Lemma 5.3** so we have the result

(End of proof) □

(22nd February, Friday)

A proof for yesterday Example Class : when showing $\sup_{Ax\leq b} c\cdot x = \min_{y\geq 0, A^T y = c} b\cdot y$. We first assume that $\exists y_0$ such that $y_0 \geq A^T y_0 = c$ (if not, can just define that the minimum takes value ∞)

Correction 1 : We defined $\Xi(x) = c\cdot x$, $\Xi : E \to \mathbb{R}$ by $A^T y_0 \cdot x = y_0 \cdot (Ax)$ so Ξ is indeed well-defined.
Correction 2 : We should have $E^* \cong \text{Range}(A) \subset \mathbb{R}^m$.

6 Existence and Characterisation of Transport Maps

Aims in this chapter are the following.

(1) To find a sufficient condition for the existence of Optimal Transport map for the MOT problem.

(2) To find a sufficient condition for the Monge cost to equal the Kantorovich cost.

(3) To characterise Optimal Transport maps and plans.

The structure would look like

6.1 Sate main results

6.2 Background on convex analysis

6.3 Prove main results

6.1 Knott-Smith Optimality and Beiner's Theorem

Definition) The **subdifferential** of a convex function φ is defined to be

$$\partial\varphi(z) = \{y : \varphi(z) \geq \varphi(x) + y \cdot (z - x) \quad \forall z \in \mathbb{R}^d\}$$

This is a *set of slopes.*

Comments :

(1) Subdifferential always exists for convex lower semi-continuous functions.

(2) if φ is differentiable at x, then $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

Theorem 6.1) *(Knott-Simith Optimality Criterion)* Let $\mu \in \mathcal{P}_2(X)$, $\nu \in \mathcal{P}_2(Y)$, $X, Y \subset \mathbb{R}^d$, $c(x,y) = \frac{1}{2}|x-y|^2$. Then $\pi^\dagger \in \Pi(\mu,\nu)$ minimises the KOT problem *iff* $\exists\tilde{\varphi}^\dagger \in L^1(\mu)$ convex and lower semi-continuous such that $\text{supp}(\pi^\dagger) \subset \text{Gra}(\partial\tilde{\varphi}^\dagger)$ (equivalent to having $y \in \partial\tilde{\varphi}^\dagger(x)$ for $\pi^\dagger$-a.e. (x,y)). Moreover the pair $(\tilde{\varphi}^\dagger, (\tilde{\varphi}^\dagger)^c)$ is a minimiser of $\inf_{\tilde{\Phi}} \mathbb{J}$, where $\tilde{\Phi} = \{(\tilde{\varphi}, \tilde{\psi}) \in L^1(\mu) \times L^1(\nu) : \tilde{\varphi}(x) + \tilde{\psi}(y) \geq x \cdot y\}$.

(Previously, $\sup_{(\varphi,\psi)\in\Phi} \mathbb{J}(\varphi,\psi)$ was the dual problem. Here we are interested in the problems $\inf_{(\tilde{\varphi},\tilde{\psi})\in\tilde{\Phi}} \mathbb{J}(\tilde{\varphi},\tilde{\psi})$. The two problems can be related by : $(\tilde{\varphi},\tilde{\psi}) \in \tilde{\Phi}$ minimises $\mathbb{J}$ over $\tilde{\Phi}$ *iff* $(\varphi,\psi) \in \Phi$ maximises $\mathbb{J}$ over Φ where $\tilde{\varphi}(x) = \frac{1}{2}|x|^2 - \varphi(x)$, $\tilde{\psi}(y) = \frac{1}{2}|y|^2 - \psi(y)$.)

In 1D we expect monotonicity, and the theorem is almost equivalent to monotonicity : if $x_1 \leq x_2$ then $T^\dagger(x_1) \leq T^\dagger(x_2)$. Hence $\text{supp}(\pi^\dagger) = \{(x,y) : x \in X, y = T(x)\}$ is **cyclically monotone** - if $(x_1,y_1),(x_2,y_2) \in \text{supp}(\pi^\dagger)$ and $x_1 \leq x_2$ then $y_1 \leq y_2$. Any cyclically monotone set can be written as a subdifferential of a convex function. This argument holds for higher dimensions and if T is "set valued".

Theorem 6.2) *(Brenier's Theorem)* Let $\mu \in \mathcal{P}_2(X)$, $\nu \in \mathcal{P}_2(Y)$, $X, Y \subset \mathbb{R}^d$ and $c(x,y) = \frac{1}{2}|x-y|^2$. Assume that μ does not give mass to small sets (a small set is any set with Hausdorff dimenstion at most $d-1$). Then there is a unique $\pi^\dagger \in \Pi(\mu,\nu)$ that minimises the KOT problem.

Moreover, $\pi^\dagger$ satisfies $\pi^\dagger = (id \times \nabla\tilde{\varphi})_{\#}\mu$ where $\nabla\tilde{\varphi}$ is the unique gradient of a convex function that pushes μ forward to ν (that is, $(\nabla\tilde{\varphi})_{\#}\mu = \nu$) and $(\tilde{\varphi}, \tilde{\varphi}^c)$ minimise $\mathbb{J}$ over $\tilde{\Phi}$.

Comments :

(1) $\pi^\dagger = (id \times \nabla\tilde{\varphi})_{\#}\mu \Leftrightarrow d\pi^\dagger(x,y) = \delta_{\nabla\tilde{\varphi}(x)}(y) \times d\mu(x)$.

(2) We will show that, in **Proposition 6.5**, convex functions are differentiable Lebesgue almost everywhere. Since μ gives zero mass to sets of Lebesgue measure 0, then any convex function is differentiable μ almost everywhere.

Corollary 6.3) Under the same assumptions as **Theorem 6.2**, $\nabla\tilde{\varphi}$ is the unique solution to the MOT problem, *i.e.*

$$\frac{1}{2}\int_X |x - \nabla\tilde{\varphi}(x)|^2 d\mu(x) = \inf_{T:T_{\#}\mu=\nu} \frac{1}{2}\int_X |x - T(x)|^2 d\mu(x)$$

proof) Since $\min \mathbb{K} \leq \inf \mathbb{M}$, and $T^\dagger_{\#}\mu = \nu$ by **Theorem 6.2**, it is enough to show that $T^\dagger = \nabla\tilde{\varphi}$ satisfies $\mathbb{M}(T^\dagger) = \min \mathbb{K} = \mathbb{K}(\pi^\dagger)$. Indeed,

$$\begin{aligned}
\mathbb{M}(T^\dagger) =& \frac{1}{2}\int_X |x - T^\dagger(x)|^2 d\mu(x) \\
=& \frac{1}{2}\int_{X\times Y} |x - T^\dagger(x)|^2 d\pi^\dagger(x,y) \\
=& \frac{1}{2}\int_{X\times Y} |x - y|^2 d\pi^\dagger(x,y) \quad \text{(since } y = T^\dagger(x),\ \pi^\dagger\text{-a.e.)} \\
=& \mathbb{K}(\pi^\dagger)
\end{aligned}$$

(End of proof) □

6.2 Preliminary Results for Convex Analysis

Just in this section, we will write φ rather than $\tilde{\varphi}$.

Recall the *Legendre-Fenchel transform*, or the convex conjugate defined by

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

Proposition 6.4) Let $\varphi : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ be proper (not identically $+\infty$), lower semi-continuous and convex function. Then $\forall x, y \in \mathbb{R}^d$,

$$xy = \varphi(x) + \varphi^*(y) \quad \Leftrightarrow \quad y \in \partial\varphi(x)$$

proof) Note, by definition, $\varphi^*(y) \geq x \cdot y - \varphi(x)$ for all $x \in \mathbb{R}^d$. So

$$\begin{aligned}
x \cdot y = \varphi(x) + \varphi^*(y) \quad &\Leftrightarrow \quad x \cdot y \geq \varphi(x) + \varphi^*(y) \\
&\Leftrightarrow \quad x \cdot y \geq \varphi(x) + y \cdot z - \varphi(z) \quad \forall z \in \mathbb{R}^d \\
&\Leftrightarrow \quad \varphi(z) \geq \varphi(x) + y \cdot (z - x) \quad \forall z \in \mathbb{R}^d \\
&\Leftrightarrow \quad y \in \partial\varphi(x)
\end{aligned}$$

(End of proof) □

Proposition 6.5) Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex, then

(1) φ is differential Lebesgue-almost everywhere on the interior of its domain and

(2) Whenever φ is differentiable, has $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

(25th February, Monday)

We use the following theorem for the proof.

Rademacher's Theorem) If $U \subset \mathbb{R}^d$ is open and $f : U \to \mathbb{R}$ is Lipschitz continuous then f is differentiable a.e.

We do not prove this results.

Proposition 6.5) Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex, then

(1) φ is differential Lebesgue-almost everywhere on the interior of its domain and

(2) Whenever φ is differentiable, has $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

proof)

(1) Let $x \in \text{int}(\text{Dom}(\varphi))$ and δ^* be such that $\overline{B(x,\delta^*)} \subset \text{int}(\text{Dom}(\varphi))$. We show that φ is Lipschitz continuous on $B(x, \delta^*/4)$. Then by *Rademacher's Theorem*, φ is a.e. differentiable in $B(x, \delta^*/4)$. Then φ is differentiable a.e. on $\text{int}(\text{Dom}(\varphi))$.

(a) We show φ is uniformly bounded on $\overline{B(x,\delta^*/2)}$. Let Q be a cuboid centred at x with sides of length $\sqrt{\frac{4}{d}}\delta^*$. Let $\{x_i\}_{i=1}^{n}$ be the corners of Q. Note $x \in \partial B(x,\delta^*)$ and that the set of extreme points of Q is $\{x_i\}_{i=1}^{d}$. by the *Minkowski-Caretheódory theorem* (Theorem 3.5), for each $y \in \overline{B(x,\delta^*/\sqrt{d})}$, $\exists\{\lambda_i\}_{i=1}^{2^d} \subset [0,1]$ such that $\sum_{i=1}^{2^n}\lambda_i = 1$ and $y = \sum_{i=1}^{2^d}\lambda_i x_i$. So

$$\varphi(y) = \varphi\Big(\sum_{i=1}^{2^d}\lambda_i x_i\Big) \le \sum_{i=1}^{2^d}\lambda_i\varphi(x_i) \le \max_i|\varphi(x_i)| = C$$

where the first inequality follows from convexity of φ. Now define $y' = (x-(y-x)) = 2x - y \in B(x,\delta^*/2)$. Since $x = \frac{1}{2}y' + \frac{1}{2}y$, has

$$\begin{aligned} &\varphi(y) \ge 2\varphi(x) - \varphi(y') \ge 2\varphi(x) - C \\ \Rightarrow\quad &2\varphi(x) - C \le \varphi(y) \le C, \quad \forall y \in \overline{B(x,\delta^*/\sqrt{d})} \end{aligned}$$

so

$$\big\|\varphi\big\|_{L^\infty(\overline{B(x,\delta^*/\sqrt{d})})} \le \max\{C - 2\varphi(x), C\} = M < \infty$$

(b) We show φ is Lipschitz continuous on $B(x,\delta^*/2\sqrt{d})$. Let $x_1, x_2 \in B(x,\delta^*/2\sqrt{d})$ where $x_1 \ne x_2$. Take x_3 to be the intersection of the line through x_1 and x_2 with $\partial B(x,\delta^*/\sqrt{d})$ and choose x_3 such that x_2 lies between x_1 and x_3. Define $\lambda = \frac{|x_2-x_3|}{|x_1-x_3|} \in (0,1)$. Now,

$$\begin{aligned} \lambda x_1 + (1-\lambda)x_3 =& \lambda x_2 + \lambda(x_1-x_2) + (1-\lambda)x_2 + (1-\lambda)(x_3-x_2) \\ =& x_2 + \frac{|x_2-x_3|}{|x_1-x_3|}(x_1-x_2) + \frac{|x_1-x_3|-|x_2-x_3|}{|x_2-x_3|}(x_3-x_2) \\ =& x_2 + \frac{1}{|x_1-x_3|}\Big(|x_2-x_3|(x_1-x_2) + |x_1-x_2|(x_3-x_2)\Big) \\ =& x_2 \end{aligned}$$

By convexity of φ,

$$\begin{aligned} &\varphi(x_2) \le (1-\lambda)\varphi(x_3) + \lambda\varphi(x_1) \\ \Rightarrow\quad &\varphi(x_2) - \varphi(x_1) \le (1-\lambda)(\varphi(x_3) - \varphi(x_1)) = \frac{|x_1-x_2|}{|x_1-x_3|}(\varphi(x_3)-\varphi(x_1)) \\ &\le \frac{2\sqrt{d}\times 2M}{\delta^*}|x_1-x_2| = L|x_1-x_2| \end{aligned}$$

with $L = 4\sqrt{d}M/\delta^*$. So φ is Lipschitz. *[This proof looks very complicated but the idea is very simple!]*

(2) Let φ be differentiable at x. Then

$$\begin{aligned}\varphi(x)+\nabla\varphi(x)\cdot(z-x)=&\varphi(x)+\lim_{h\to 0^+}\left(\frac{\varphi(x+(z-x)h)-\varphi(x)}{h}\right)\\ =&\varphi(x)+\lim_{h\to 0^+}\left(\frac{\varphi((1-h)x+zh)-\varphi(x)}{h}\right)\\ \leq&\varphi(x)+\lim_{h\to 0}\frac{(1-h)\varphi(x)+h\varphi(z)-\varphi(x)}{h}=\varphi(z)\end{aligned}$$

so $\nabla\varphi(x)\in\partial\varphi(x)$.

On the other hand, if $y\in\partial\varphi(x)$, then $\varphi(x)+y\cdot(z-x)\leq\varphi(x)$ for all $z\in\mathbb{R}^d$ so by letting $z=x+hw$ with $h>0$, we get

$$y\cdot w\leq\frac{\varphi(x+hw)-\varphi(x)}{h}$$

Let $h\to 0^+$, then $y\cdot w\leq\nabla\varphi(x)\cdot w$. By symmetry ($w\mapsto -w$), we also have $y\cdot w=\nabla\varphi(x)\cdot w$ for all $w\in\mathbb{R}^d$. Hence $y=\nabla\varphi(x)$.

(End of proof) □

Proposition 6.6) Let $\varphi:\mathbb{R}\cup\{+\infty\}$ be proper. then the following are equivalent.

(1) φ is convex and lower semi-continuous.

(2) $\varphi=\psi^*$ for some proper function ψ.

(3) $\varphi^{**}=\varphi$.

proof) The implications from (3) to (2) is immediate.

We are left to show implication (2) $\Rightarrow$ (1). Assume (2), so $\varphi=\psi^*$ for some proper function ψ. Let $x_1,x_2\in\mathbb{R}^d$, $t\in[0,1]$, then

$$\begin{aligned}\varphi(tx_1+(1-t)x_2)&=\sup_y\Big((tx_1+(1-t)x_2)\cdot y-\psi(y)\Big)\\ &\leq\sup_y\Big(t(x_1\cdot y)-\psi(y)\Big)+\sup_y\Big((1-t)(x_2\cdot y-\psi(y)\Big)\\ &=t\varphi(x_1)+(1-t)\varphi(x_2)\end{aligned}$$

so φ is convex. To show lower semi-continuity, let $x_m\to x$. Then

$$\begin{aligned}\liminf_{m\to\infty}\varphi(x_m)=&\liminf_{m\to\infty}\sup_y(x_m\cdot y-\psi(y))\geq\lim_{m\to\infty}(x_m\cdot y-\psi(y))\\ =&x\cdot y-\psi(y)\quad\text{for any } y\in\mathbb{R}^d\end{aligned}$$

so $\liminf_{m\to\infty}\varphi(x_m)\geq\sup_y(x\cdot y-\psi(y))=\varphi(x)$, so φ is lower semi-continuous.

(27th February, Wednesday)

proof continued) Next we prove implication (1) to (3). Suppose φ is convex and lower semi-continuous. Assume $x \in \mathrm{int}(\mathrm{Dom}(\varphi))$ and we show $\varphi^{**}(x) = \varphi(x)$. Since φ can be bounded below by an affine function passing through $(x, \varphi(x))$, we have $\partial\varphi(x) \neq \phi$. Let $y_0 \in \partial\varphi(x)$. By **Proposition 6.4**, $x \cdot y_0 = \varphi(x) + \varphi^*(y_0)$, so

$$\varphi(x) = x \cdot y_0 - \varphi^*(y_0) \leq \sup_{y \in \mathbb{R}^n} (x \cdot y - \varphi^c(y)) = \varphi^{**}(x)$$

On the other hand, since $\varphi^*(y) \geq x \cdot y - \varphi(x)$ for all $y \in \mathbb{R}^d$, has

$$\varphi(x) \geq \sup_y (x \cdot y - \varphi^*(y)) = \varphi^{**}(x)$$

so $\varphi(x) = \varphi^{**}(x)$ on $\mathrm{int}(\mathrm{Dom}(\varphi))$. See note for the case $x \notin \mathrm{int}(\mathrm{Dom}(\varphi))$.

(End of proof) □

6.3 Proof of the Knott-Smith Optimality Criterion

Let $c(x,y) = \frac{1}{2}|x-y|^2$, $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. We first make few observations.

(A) Let $(\varphi, \psi) \in \Phi_c$. Define $\tilde{\varphi}(x) = \frac{1}{2}|x|^2 - \varphi(x)$ and $\tilde{\psi}(y) = \frac{1}{2}|y|^2 - \psi(y)$. One can see that $\tilde{\varphi} \in L^1(\mu)$, $\tilde{\psi} \in L^1(\nu)$. and

$$\begin{aligned}\tilde{\varphi}(x) + \tilde{\psi}(y) =& \frac{1}{2}|x|^2 + \frac{1}{2}|y|^2 - \varphi(x) - \varphi(y) \\ \geq& \frac{1}{2}|x|^2 + \frac{1}{2}|y|^2 - \frac{1}{2}|x-y|^2 = x \cdot y\end{aligned}$$

So $(\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi}$ where $\tilde{\Phi} = \{(\tilde{\varphi}, \tilde{\psi}) \in L^1(\mu) \times L^1(\nu) : \tilde{\varphi}(x) + \tilde{\psi}(y) \geq x \cdot y\}$.

Similarly if $(\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi}$, then $(\varphi, \psi) \in \Phi_c$.

(B) If we let $M = \frac{1}{2}\int_X |x|^2 d\mu(x) + \frac{1}{2}\int_Y |y|^2 d\nu(y)$, then $\mathbb{J}(\tilde{\varphi}, \tilde{\psi}) = M - \mathbb{J}(\varphi, \psi)$ and for $\pi \in \Pi(\mu, \nu)$, has $\mathbb{K}(\pi) = M - \int_{X \times Y} x \cdot y d\pi(x,y)$. Kantorovich duality (**Theorem 4.1**) implies that

$$\min_{(\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi}} \mathbb{J}(\tilde{\varphi}, \tilde{\psi}) = \max_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} x \cdot y d\pi(x,y)$$

Also

$$\pi^\dagger \in \Pi(\mu,\nu) \text{ minimises } \mathbb{K} \text{ over } \Pi(\mu,\nu) \;\Leftrightarrow\; \pi^\dagger \text{ maximises } \int_{X \times Y} x \cdot y d\pi(x,y)$$
$$(\varphi, \psi) \in \Phi_c \text{ maximises } \mathbb{J} \text{ over } \Phi_c \;\Leftrightarrow\; (\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi} \text{ minimise } \mathbb{J} \text{ over } \tilde{\Phi}.$$

(C) Recall that there exists maximiser $(\varphi^{cc}, \varphi^c) \in \Phi_c$ of $\mathbb{J}$. So $(\tilde{\varphi}, \tilde{\psi}) := (\frac{1}{2}|\cdot|^2 - \varphi^{cc}, \frac{1}{2}|\cdot|^2 - \varphi^c)$ minimise $\mathbb{J}$ over $\tilde{\Phi}$. Furthermore

$$\begin{aligned}\tilde{\psi}(y) =& \frac{1}{2}|y|^2 - \varphi^c(y) = \sup_{x \in X}\left(\frac{1}{2}|y|^2 - \frac{1}{2}|x-y|^2 + \varphi(y)\right) \\ =& \sup_{x \in X}\left(x \cdot y - \frac{1}{2}|x|^2 + \varphi(x)\right) = \sup_{x \in X}(x \cdot y - \tilde{\varphi}(x)) = \tilde{\varphi}^*(y)\end{aligned}$$

And

$$\begin{aligned}\tilde{\varphi}(x) =&\frac{1}{2}|x|^2 - \varphi^{cc}(x) = \sup_{y\in Y}\Big(\frac{1}{2}|x|^2 - \frac{1}{2}|x-y|^2 + \varphi^c(y)\Big)\\ =&\sup_{y\in Y}\Big(\frac{1}{2}|x|^2 - \frac{1}{2}|x-y|^2 + \frac{1}{2}|y|^2 - \tilde{\varphi}^*(y)\Big) \quad \text{(used previous computation)}\\ =&\sup_{y\in Y}(x\cdot y - \tilde{\varphi}^*(y))\\ =&\tilde{\varphi}^{**}(x)\end{aligned}$$

By **Proposition 6.6**, $\tilde{\eta} := \tilde{\varphi}^{**}$ is convex and lower semi-continuous and $\tilde{\eta}^* = \tilde{\varphi}^{***} = \tilde{\varphi}^*$.

(D) For $(\tilde{\varphi}, \tilde{\varphi}^*)$ with $\tilde{\varphi} \in L^1(\mu)$, we have

$$\begin{aligned}\int_{X\times Y}\tilde{\varphi}(x) + \tilde{\varphi}^*(y)d\pi^\dagger(x,y) &= \int_{X\times Y} x\cdot y d\pi^\dagger(x,y)\\ &\leq \frac{1}{2}\int_{X\times Y}|x|^2 + |y|^2 d\pi^\dagger(x,) = \frac{1}{2}\int_X |x|^2 d\mu(x) + \frac{1}{2}\int_Y |y|^2 d\nu(y)\end{aligned}$$

so $\mathbb{J}(\tilde{\varphi}, \tilde{\varphi}^*) < \infty$.

(E) From the result of (C), if we just prove that $\tilde{\varphi}^* \in L^1(\nu)$ whenever $\tilde{\varphi} \in L^1(\mu)$, then $(\eta, \eta^*) \in L^1(\mu) \times L^1(\nu)$ so there is a minimiser of $\mathbb{J}$ and $\tilde{\Phi}$ that takes the form $(\tilde{\eta}, \tilde{\eta}^*)$ where $\tilde{\eta}$ is convex, lower semi-continuous and is proper.

To see this, assume $\tilde{\varphi} \in L^1(\mu)$. First note that $\exists x_0 \in X$ and $b_0 = \tilde{\varphi}(x_0) + 1 \in \mathbb{R}$ such that

$$\tilde{\varphi}^*(y) \geq x_0 \cdot y - \tilde{\varphi}(x_0) - 1 =: x_0\cdot y - b_0 =: f(y)$$

Then we have $\tilde{\varphi}^* - f(y) \geq 0$, so $\Big\|\tilde{\varphi}^* - f\Big\|_{L^1(\mu)} = \int_Y \Big(\tilde{\varphi}^*(y) - f(y)\Big)d\nu(y)$. Hence

$$\begin{aligned}\Big\|\tilde{\varphi}^* - f\Big\|_{L^1(\mu)} =& \int_Y\Big(\tilde{\varphi}^*(y) - f(y)\Big)d\nu(y)\\ \leq&\mathbb{J}(\tilde{\varphi}, \tilde{\varphi}^*) + \Big\|\tilde{\varphi}\Big\|_{L^1(\mu)} + \frac{1}{2}|x_0| + \frac{1}{2}\int_Y |y|^2 d\nu(y) + b_0\\ <&+\infty\end{aligned}$$

where the second line is implied by Cauchy-Schwarz inequality. So $\tilde{\varphi}^* - f \in L^1(\nu)$ and since $f \in L^1(\nu)$, we conclude $\tilde{\varphi}^* \in L^1(\nu)$ as required.

We now come back to the one of the two main theorems of the chapter.

Theorem 6.1) *(Knott-Smith(KS) optimality criterion)* Let $\mu \in \mathcal{P}_2(X)$, $\nu \in \mathcal{P}_2(Y)$, $X, Y \in \mathbb{R}^d$, $c(x,y) = \frac{1}{2}|x-y|^2$. Then $\pi^\dagger \in \Pi(\mu,\nu)$ minimises $\mathbb{K}$ over $\Pi(\mu,\nu)$ *iff* there exists $\tilde{\varphi} \in L^1(\mu)$ convex, lower-semicontinuous such that $y \in \partial\tilde{\varphi}(x)$ for $\pi^\dagger$-a.e. (x,y).
Moreover $(\tilde{\varphi}, \tilde{\varphi}^*)$ maximises $\mathbb{J}$ over $\tilde{\Phi}$.

proof) Let $\pi^\dagger \in \Pi(\mu,\nu)$ minimise $\mathbb{K}$ over $\Pi(\mu,\nu)$ and $(\tilde{\varphi}, \tilde{\varphi}^*) \in \tilde{\Phi}$ minimise $\mathbb{J}$ over $\tilde{\Phi}$. By *Kantorovich duality*,

$$\begin{aligned}&\int_X \tilde{\varphi}(x)d\mu(x) + \int_Y \tilde{\varphi}^*(y)d\nu(y) = \int_{X\times Y} x\cdot y d\pi^\dagger(x,y)\\ \Rightarrow\quad &\int_{X\times Y}\Big(\tilde{\varphi}(x) + \tilde{\varphi}^*(y) - x\cdot y\Big)d\pi^\dagger(x,y) = 0\\ \Rightarrow\quad &\tilde{\varphi}(x) + \tilde{\varphi}^*(y) = x\cdot y \quad \text{for } \pi^\dagger\text{-a.e. } (x,y) \quad \text{(by property of } \tilde{Phi})\\ \Rightarrow\quad &y \in \partial\tilde{\varphi}(x) \quad \text{for } \pi^\dagger\text{-a.e. } (x,y)\end{aligned}$$

Conversely suppose $\pi^\dagger \in \Pi(\mu,\nu)$ and $\tilde{\varphi} \in L^1(\mu)$ and satisfy $y \in \partial\tilde{\varphi}(x)$ for $\pi^\dagger$-a.e. (x,y) where $\tilde{\varphi}$ is lower semi-continuous and convex. We want to show that $\pi^\dagger$ is optimal for KOT problem and $(\tilde{\varphi},\tilde{\varphi}^*)$ optimal for "KD" problem.

By **Proposition 6.4**,

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

By point (D) above, we have $\tilde{\varphi}^* \in L^1(\nu)$, so $(\tilde{\varphi},\tilde{\varphi}^*) \in \tilde{\Phi}$. Then

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

By duality, LHS=RHS, so all inequalities above are in fact equalities. So $(\tilde{\varphi},\tilde{\varphi}^*)$ minimise $\mathbb{J}$ over $\tilde{\Phi}$ and $\pi^\dagger$ maximises $\int_{X\times Y} x\cdot y d\pi(x,y)$ over $\Pi(\mu,\nu)$.

(End of proof) □

6.4 Proof of Brenier's Theorem

In this section, we prove the second one of the two main theorems of the chapter.

Theorem 6.2) *(Brenier's Theorem)* Let $\mu \in \mathcal{P}_2(X)$, $\nu \in \mathcal{P}_2(Y)$, $X,Y \subset \mathbb{R}^d$ and $c(x,y) = \frac{1}{2}|x-y|^2$. Assume that μ does not give mass to small sets (a small set is any set with Hausdorff dimension at most $d-1$). Then there is a unique $\pi^\dagger \in \Pi(\mu,\nu)$ that minimises the KOT problem.

Moreover, $\pi^\dagger$ satisfies $\pi^\dagger = (id\times\nabla\tilde{\varphi})_\#\mu$ where $\nabla\tilde{\varphi}$ is the unique gradient of a convex function that pushes μ forward to ν (that is, $(\nabla\tilde{\varphi})_\#\mu = \nu$) and $(\tilde{\varphi},\tilde{\varphi}^c)$ minimise $\mathbb{J}$ over $\tilde{\Phi}$.

(1st March, Friday)

(We have a lecture on the 15th March)
(Exercies Class 3 is at Tuesday, 5th March, 4pm, MR13)

proof of Brenier's theorem) Let $\pi^\dagger \in \Pi(\mu,\nu)$ minimise KOT problem and $\{\pi^\dagger(\cdot|x)\}_{x\in X}$ be the disintegration of measure,

$$\pi^\dagger(A\times B) = \int_A \pi^\dagger(B|x) d\mu(x)$$

By **Theorem 6.1**, for μ-a.e. $x \in X$ and $\pi^\dagger(\cdot|x)$-a.e. $y \in Y$ we have $y \in \partial\tilde{\varphi}(x)$ where $\tilde{\varphi}$ minimises $\mathbb{J}$ over $\tilde{\Phi}$. Since $\partial\tilde{\varphi}(x) = \{\nabla\tilde{\varphi}(x)\}$ for μ-a.e. $x \in X$, has $y = \nabla\tilde{\varphi}(x)$ for μ-a.e. $x \in X$ and $\pi^\dagger(\cdot|x)$-a.e. $y \in Y$. Hence $\pi^\dagger(\cdot|x) = \delta_{\nabla\tilde{\varphi}(x)}$. So

$$\begin{aligned}\pi^\dagger(A\times B) = \int_A \mathbf{1}_{\nabla\tilde{\varphi}(x)\in B} d\mu(x) =&\mu\Big(\{x : x\in A \text{ and } \nabla\tilde{\varphi}(x)\in B\}\Big)\\ =&(\mathrm{Id}\times\nabla\tilde{\varphi})_\#\mu(A\times B)\end{aligned}$$

so $\pi^\dagger = (Id\times\nabla\tilde{\varphi})_\#\mu$. Also

$$\begin{aligned}\nu(B) = \pi^\dagger(X\times B) &= (\mathrm{Id}\times\nabla\tilde{\varphi})_\#\mu(X\times B)\\ &= \mu\Big(\{x : (x,\nabla\tilde{\varphi}(x)) \in X\times B\}\Big)\\ &= \mu\Big(\{x : \nabla\tilde{\varphi}(x)\in B\}\Big) = (\nabla\tilde{\varphi})_\#\mu(B)\end{aligned}$$

For uniqueness, suppose $\overline{\varphi}$ is convex and satisfies $(\nabla\overline{\varphi})_{\#}\mu = \nu$. We show $\nabla\tilde{\varphi} = \nabla\overline{\varphi}$ a.e. By **Theorem 6.1**,

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

is an optimal transport plan and $(\overline{\varphi}, \overline{\varphi}^c)$ minimise $\mathbb{J}$ over $\tilde{\Phi}$. So

$$\begin{aligned}
& \int_X \overline{\varphi} d\mu + \int_Y \overline{\varphi}^c d\nu = \int_X \tilde{\varphi} d\mu + \int_Y \tilde{\varphi}^c d\nu \\
\Rightarrow \quad & \int_{X\times Y} (\overline{\varphi}(x) + \overline{\varphi}^c(y)) d\pi^\dagger(x,y) = \int_{X\times Y} (\tilde{\varphi}(x) + \tilde{\varphi}^c(y)) d\pi^\dagger(x,y) \\
& = \int_{X\times Y} x\cdot y \, d\pi^\dagger(x,y) \quad \text{by } \textbf{Proposition 6.4} \\
& = \int_{X\times Y} x\cdot y \, d(\mathrm{Id}\times\nabla\tilde{\varphi})_{\#}\mu(x,y) \\
& = \int_X x\cdot\nabla\tilde{\varphi}(x) d\mu(x)
\end{aligned}$$

Also,

$$\int_{X\times Y} \Big(\overline{\varphi}(x) + \overline{\varphi}^*(y)\Big) d\pi^\dagger(x,y) = \int_X \Big(\overline{\varphi}(x) + \overline{\varphi}^*(\nabla\tilde{\varphi}(x))\Big) d\mu(x)$$

so

$$\int_X \Big(\overline{\varphi}(x) + \overline{\varphi}^*(\nabla\tilde{\varphi}(x)) - x\cdot\nabla\tilde{\varphi}(x)\Big) d\mu(x) = 0$$

but $\overline{\varphi}(x) + \overline{\varphi}^*(y) \geq x\cdot y$, so

$$\overline{\varphi}(x) + \overline{\varphi}^*(\nabla\widetilde{\varphi(x})) - x\cdot\nabla\tilde{\varphi}(x) = 0 \quad \mu\text{-a.e. } x\in X$$

By **Proposition 6.4**, $\nabla\tilde{\varphi}(x) \in \partial\overline{\varphi}(x) = \{\nabla\overline{\varphi}(x)\}$, so by **Propositioin 6.5**, $\nabla\tilde{\varphi}(x) = \nabla\overline{\varphi}(x)$ for μ-a.e. $x \in X$.

(End of proof) □

7 Wasserstein Distances

In this chapter, we assume $c(x,y) = |x-y|^p$ with $p \in [1,+\infty)$ with $X = Y \subset \mathbb{R}^d$.

The objectives in this chapter are

(1) Define the *Wasserstein distance* d_{W^p} and show it is a metric in $\mathcal{P}_p(X)$.

(2) Show equivalence of d_{W^p} and d_{W^q} when X is bounded.

(3) Relationship with weak-* topology.

(4) Show that $(\mathcal{P}_p(X), d_{W^p})$ is a *geodesic space* (to be defined later).

Another interesting cost function that does not fit into the Wasserstein framework is $c(x,y) = 1_{x\neq y}$.

Proposition 7.1) Let $\mu,\nu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, $c(x,y) = \mathbf{1}_{x\neq y}$. Then

$$\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi) = \frac{1}{2}\|\mu-\nu\|_{TV}$$

where $\|\mu\|_{TV} = 2\sup_A |\mu(A)|$.

proof) By the *KR theorem* (**Theorem 4.13**),

$$\min_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi) = \sup\Big\{ \int_X f d(\mu-\nu) : f \in L^1(|\mu-\nu|), \big\|f\big\|_{Lip} \leq 1\Big\}$$
$$= \sup\Big\{ \int_X f d(\mu-\nu) : 0 \leq f(x) \leq 1 \ \forall x \in X\Big\}$$

where $\|f\|_{Lip}$ is given in terms of c rather than a usual metric. Write $\mu-\nu = (\mu-\nu)_+ - (\mu-\nu)_-$ where $(\mu-\nu)_\pm \in \mathcal{M}_+(X)$ and are singular (*i.e.* has disjoint support). We may achieve the supremum when we choose $f(x) = 1$ if $x \in \operatorname{supp}(\mu-\nu)_+$ and $f(x) = 0$ otherwise. Then

$$\min_{\pi\in\Pi(\mu,\nu)} \mathbb{K}(\pi) = (\mu-\nu)_+(X)$$

For TV, the optimal choice is $A = \operatorname{supp}((\mu-\nu)_+)$ and $\|\mu-\nu\|_{TV} = 2(\mu-\nu)_+(X)$. So $\min\mathbb{K} = \frac{1}{2}\|\mu-\nu\|_{TV}$.

(End of proof) □

7.1 Wasserstein Distances

We work on the space of measures with bounded p^{th} moment,

$$\mathcal{P}_p(X) = \{\mu \in \mathcal{P}(X) : \int_X |x|^p d\mu(x) < +\infty\}$$

If X is bounded, then $\mathcal{P}_p(X) = \mathcal{P}(X)$.

Definition 7.2) Let $\mu,\nu \in \mathcal{P}_p(X)$, then the p **-Wasserstien distance** is defined to be

$$d_{W^p}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \Big(\int_{X\times Y} |x-y|^p d\pi(x,y)\Big)^{1/p}$$

If $\mu \in \mathcal{P}_p(X)$, $\nu \in \mathcal{P}_p(X)$, then

$$d_{W^p}(\mu,\nu)^p = \min_{\pi\in\Pi(\mu,\nu)} p\int_{X\times Y} |x|^p + |y|^p d\pi(x,y) = p\int |x|^p d\mu(x) + p\int |y|^p d\nu(y) < +\infty$$

(4th March, Monday)

Recall, we defined

Definition 7.2) For $\mu,\nu \in \mathcal{P}_p(X)$, the **Wasserstein distance** is defined by

$$d_{W^p}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \Big(\int_{X\times Y} |x-y|^p d\pi(x,y)\Big)^{1/p}$$

Proposition 7.3) Let $X \subset \mathbb{R}^d$. Then the distance $d_{W^p} : \mathcal{P}_p(X)\times\mathcal{P}_p(X) \to [0,\infty)$ is a metric.

proof)

(1) It is easy to see that $d_{W^p}(\mu,\nu) \geq 0$ for any μ, ν.

(2) If $\mu = \nu$, then $\pi(x, y) = \delta_y(x)\mu(x)$ satisfies $\pi \in \Pi(\mu, \nu)$ and

$$d_{W^p}(\mu, \nu) \le \int_{X\times Y} |x - y|^p d\pi(x, y) = 0$$

If $d_{W^p}(\mu, \nu) = 0$, then $\exists \pi \in \Pi(\mu, \nu)$ such that $\int_{X\times Y} |x - y|^p d\pi(x, y) = 0$ so $x = y$ for π-a.e. x, y. So for any function $f : X \to \mathbb{R}$,

$$\int_X f(x)d\mu(x) = \int_{X\times X} f(x)d\pi(x, y) = \int_{X\times X} f(y)d\pi(x, y) = \int_X f(y)d\nu(y) \quad \cdots\cdots\cdots \ (*)$$

Since $(*)$ holds for all $f : X \to \mathbb{R}$ integrable w.r.t μ and ν (or just for f continuous and bounded is sufficient), we have $\mu = \nu$.

(3) Clearly $d_{W^p}(\mu, \nu) = d_{W^p}(\nu, \mu)$. (To write out formally, define $s : (x, y) \mapsto (y, x)$ and observe that whenever $\pi \in \Pi(\mu, \nu)$, has $s_{\#}\Pi(\nu, \mu)$)

(4) To see the triangular inequality, we make use of *glueing lemma* (**Lemma 7.4**). Let $\mu, \nu, \omega \in \mathcal{P}_p(X)$ and $\pi_{XY} \in \pi(\mu, \nu)$ and $\pi_{YZ} \in \pi(\nu, \omega)$ be the optimal plans, *i.e.*

$$\begin{aligned} d_{W^p}(\mu, \nu) &= \Big(\int_{X\times X} |x - y|^p d\pi_{XY}(x, y) \Big)^{1/p} \\ d_{W^p}(\nu, \omega) &= \Big(\int_{X\times X} |y - z|^p d\pi_{YZ}(y, z) \Big)^{1/p} \end{aligned}$$

Take $\gamma \in \mathcal{P}(X \times Y \times Z)$ such that $P^{X,Y}_{\#}\gamma = \pi_{X\times Y}$ and $P^{Y}_{\#}\gamma = \pi_{YZ}$ using the *glueing lemma*. Define $\pi_{XZ} = P^{X,Z}_{\#}\gamma$. Since $\pi_{XZ}(A \times X) = \gamma(A \times X \times X) = \pi_{XY}(A\times X) = \mu(A)$ and similarly $\pi_{XZ}(X\times C) = \omega(C)$, we see that $\pi_{XZ} \in \Pi(\mu, \omega)$. Now, by *Minkowski's inequality,*

$$\begin{aligned} d_{W^p}(\mu, \omega) &\le \Big(\int_{X\times X} |x - z|^p d\pi_{XZ}(x, z) \Big)^{1/p} = \Big(\int_{X\times X\times X} |x - z|^p d\gamma(x, y, z) \Big)^{1/p} \\ &\le \Big(\int_{X\times X\times X} |x - y|^p d\gamma(x, y, z) \Big)^{1/p} + \Big(\int_{X\times X\times X} |y - z|^p d\gamma(x, y, z) \Big)^{1/p} \\ &= \Big(\int_{X\times X} |x - y|^p d\pi_{XY}(x, y) \Big)^{1/p} + \Big(\int_{X\times X} |y - z|^p d\pi_{YZ}(y, z) \Big)^{1/p} \\ &= d_{W^p}(\mu, \nu) + d_{W^p}(\nu, \omega) \end{aligned}$$

(End of proof) □

For the triangular inequality, we needed the following *glueing lemma.*

Lemma 7.4) Let $X, Y, Z \subset \mathbb{R}^d$, $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$, $\omega \in \mathcal{P}(Z)$, $\pi_1 \in \Pi(\mu, \nu)$, $\pi_2 \in \Pi(\nu, \omega)$. Then there is a measure $\gamma \in \mathcal{P}(X \times Y \times Z)$ such that $P^{X\times Y}_{\#}\gamma = \pi_1$ and $P^{Y,Z}_{\#} = \pi_{12}$ where $P^{X,Y}(x, y, z) = (x, y)$, $P^{Y,Z}(x, y, z) = (y, z)$.

proof) By disintegration of measures, we can write

$$\pi_1(A \times B) = \int_B \pi_1(A|y)d\nu(y), \quad \pi_2(B \times C) = \int_B \pi_2(C|y)d\nu(y)$$

for families $\{\pi_1(\cdot|y)\}_{y\in Y} \subset \mathcal{P}(X)$ and $\{\pi_2(\cdot|y)\}_{y\in Y} \subset \mathcal{P}(Z)$. Define $\gamma \in \mathcal{M}(X \times Y \times Z)$ by

$$\gamma(A \times B \times C) = \int_B \pi_1(A|y)\pi_2(C|y)d\nu(y)$$

We can check

$$\gamma(A \times B \times Z) = \int_B \pi_1(A|y)\pi_2(Z|y)d\nu(y) = \pi_1(A \times B)$$
$$\gamma(X \times B \times C) = \pi_2(B \times C)$$

So $P^{X,Y}_{\#}(\gamma) = \pi_1$ and $P^{Y,Z}_{\#}\gamma = \pi_2$. It also follows that $\gamma \in \mathcal{P}(X \times Y \times Z)$.

(End of proof) □

Proposition 7.5) Let $X \subset \mathbb{R}^d$. For every $p, q \in [1, +\infty)$, $q \leq p$ and any $\mu, \nu \in \mathcal{P}_p(X)$, we have $d_{W^p}(\mu, \nu) \geq d_{W^q}(\mu, \nu)$.

Furthermore, if X is bounded then $d^p_{W^p}(\mu, \nu) \leq \operatorname{diam}(X)^{p-1} d_{W^1}(\mu, \nu)$ (where $\operatorname{diam}(X) = \sup_{w,z \in X} |w - z|$)

proof) The first part uses Jensen's inequality. The second part makes use of Hölder inequality.

(End of proof) □

7.2 The Wasserstein Topology

Aim : show $\mu_n \xrightarrow{w-*} \mu$ *iff* $d_{W^p}(\mu_n, \mu) \to 0$. We start with the case when $X \subset \mathbb{R}^d$ is compact.

Theorem 7.6) Let $X \subset \mathbb{R}^d$ be *compact*, $\{\mu_n\}_{n \in \mathbb{N}} \subset \mathcal{P}(X)$ and $\mu \in \mathcal{P}(X)$ and $p \in [1, \infty)$. Then

$$\mu_m \xrightarrow{w-*} \mu \quad \textit{iff} \quad d_{W^p}(\mu_m, \mu) \to 0$$

proof) By **Proposition 7.5**, it is enough to prove the result for $p = 1$.

Assume $d_{W^1}(\mu_m, \mu) \to 0$. by the *Kantorovich-Rubinstein Theorem* (**Theorem 4.13**),

$$d_{W^1}(\mu, \nu) = \sup\left\{ \int_X \varphi d(\mu - \nu) : \varphi \in L^1(|\mu - \nu|), |\varphi(x) - \varphi(y)| \leq |x - y| \right\}$$

Let φ be Lipschitz with $\|\varphi\|_{Lip} > 0$. Then define $\tilde{\varphi} = \frac{1}{\|\varphi\|_{Lip}}\varphi$ so that $\tilde{\varphi}$ is 1-Lipschitz. So

$$\frac{1}{\|\varphi\|_{Lip}} \int \varphi d(\mu_m - \mu) = \int \tilde{\varphi}(\mu_m - \mu) \leq d_{W^1}(\mu_m, \mu) \to 0$$

Then $\limsup_{m\to\infty} \int \varphi d\mu_m \leq \int \varphi d\mu$. Substituting $\varphi \mapsto -\varphi$ gives the inverse inequality, so together we have

$$\lim_{m\to\infty} \int_X \varphi d\mu_n = \int \varphi d\mu$$

By *Portmanteau Theorem* (**Theorem 1.2**), has $\mu_m \xrightarrow{w-*} \mu$.

Conversely, assume $\mu_n \xrightarrow{w-*} \mu$ and let $(m_k)_k \subset \mathbb{N}$ be the subsequence such that

$$\lim_{k\to\infty} d_{W^1}(\mu_{m_k}, \mu) = \limsup_{m\to\infty} d_{W^1}(\mu_m, \mu)$$

Let $\tilde{\varphi}_{m_k}$ be 1-Lipschitz and such that $d_{W^1}(\mu_{m_k}, \mu) \leq \int_X \tilde{\varphi}_{m_k} d(\mu_{m_k} - \mu) + \frac{1}{k}$. Pick $x_0 \in \operatorname{supp}(\mu)$ and define $\varphi_{m_k}(x) = \tilde{\varphi}_{m_k}(x) - \tilde{\varphi}_{m_k}(x_0)$. So,

$$\begin{aligned} d_{W^1}(\mu_{m_k}, \mu) &\leq \int_X \varphi_{m_k} d(\mu_{m_k} - \mu) - \tilde{\varphi}_{m_k}(x_0) \int_X d(\mu_{m_k} - \mu) + \frac{1}{k} \\ &= \int_X \varphi_{m_k} d(\mu_{m_k} - \mu) + \frac{1}{k} \end{aligned}$$

Since φ_{m_k} are 1-Lipschitz and $\varphi_{m_k}(x_0) = 0$, we have $\{\varphi_{m_k}\}_{k \in \mathbb{N}}$ uniformly bounded and equi-continuous. By the *Arzelà-Ascoli Theorem*, there is a subsequence (relabelled) such that $\varphi_{m_k} \to \varphi$ uniformly and φ is 1-Lipschitz. Hence,

$$\begin{aligned} \limsup_{m \to \infty} d_{W^1}(\mu_m, \mu) &\leq \limsup_{k \to \infty} \left(\int_X \varphi_{m_k} d(\mu_{m_k} - \mu) + \frac{1}{k} \right) \\ &= \limsup_{k \to \infty} (\int_X (\varphi_{m_k} - \varphi) d(\mu_{m_k} - \mu) + \int_X \varphi d(\mu_{m_k} - \mu)) \\ &\leq \limsup_{k \to \infty} \left\| \varphi_{m_k} - \varphi \right\|_\infty + \limsup_{k \to \infty} \int_X \varphi d(\mu_{m_k} - \mu) = 0 \end{aligned}$$

(End of proof) □

(6th March, Wednesday)

Theorem 7.7) $\mu_n, \mu \in \mathcal{P}_p(\mathbb{R}^d)$. Then

$$d_{W^p}(\mu_n, \mu) \to 0 \quad \textit{iff} \quad \mu_n \xrightarrow{w^*} \mu \text{ and } \int_{\mathbb{R}^d} |x|^p d\mu_n \to \int_{\mathbb{R}^d} |x|^p d\mu$$

proof) Let $d_{W^p}(\mu_n, \mu) \to 0$. By **Proposition 7.5**, $d_{W^1}(\mu_n, \mu) \to 0$. Analogous to the proof of **Theorem 7.6**, we have $\int \varphi d(\mu_n - \mu) \to 0$ for all Lipschitz functions φ. So $\mu_n \xrightarrow{w^*} \mu$.

To show $\int |x|^p d\mu_n(x) \to \int |x|^p d\mu(x)$, we write

$$\int_{\mathbb{R}^d} |x|^p d\mu_n(x) = d^p_{W^p}(\mu_n, \delta_0), \quad \int_{\mathbb{R}^d} |x|^0 d\mu(x) = d^p_{W^p}(\mu, \delta_0)$$

Also by triangular inequality,

$$d_{W^p}(\mu, \delta_0) - d_{W^p}(\mu, \mu_n) \leq d_{W^p}(\mu_n, \delta_0) = \left(\int |x|^p d\mu_n(x) \right)^{1/p} \leq d_{W^p}(\mu_n, \mu) + d_{W^p}(\mu, \delta_0) \to d_{W^p}(\mu, \delta_0)$$

but both very LHS and the RHS converges to $d_{W^p}(\mu, \delta_0)$, so $d_{W^p}(\mu_n, \delta_0) \to d_{W^p}(\mu, \delta_0)$ as desired.

To show the converse, let $\mu_n \xrightarrow{w^*} \mu$ and $\int |x|^p d\mu \to \int |x|^p d\mu$. For any $R > 0$, define $\phi_R(x) = (\min\{|x|, R\})^p$ which is continuous and bounded. So

$$\int_{\mathbb{R}^d} |x|^p - \phi_R(x) d\mu_n \to \int_{\mathbb{R}^d} |x|^p - \phi_R(x) d\mu \quad \cdots\cdots\cdots \ (*)$$

Now

$$\int_{\mathbb{R}^d} (|x|^p - \phi_R(x)) d\mu(x) = \int_{|x|>R} (|x|^p - R^p) d\mu \leq \int_{|x|>R} |x|^p d\mu$$

For any $\epsilon > 0$, we can find $R > 0$ such that $\int_{\mathbb{R}^d}(|x|^p - \phi_R(x))d\mu(x) < \epsilon/2$. By $(*)$, for m sufficiently large, $\int_{\mathbb{R}^d}(|x|^p - \phi_R(x))d\mu_m(x) < \epsilon$. Since $(a+b)^p \geq (a^p + b^p)$ for any $a, b \geq 0$, for $|x| > R$, we have $(|x| - R)^p \leq |x|^p - R^p = |x|^p - \phi_R(x)$. So

$$\int_{|x|>R}(|x| - R)^p d\mu_m < \epsilon, \quad \int_{|x|>R}(|x| - R)^p d\mu < \epsilon$$

Let $P_R : \mathbb{R}^d \to \overline{B(0,R)}$ be the projection onto $\overline{B(0,R)}$, *i.e.* $P_R = x$ if $x \in \overline{B(0,R)}$ and $P_R(x) = xR/|x|$ is otherwise.Then P_R is continuous, and $P_R = id$ on $\overline{B(0,R)}$, and for $x \notin \overline{B(0,R)}$, we have $|x - P_R(x)| = |x| - R$. Hence

$$\begin{aligned} d_{W^p}(\mu, (P_R)_{\#}\mu) &\leq \Big(\int_{\mathbb{R}^d} |x - P_R(x)|^p d\mu(x)\Big)^{1/p} \quad \text{since } d_{W^p}(\cdot,\cdot) \text{ is optimal for MOT} \\ &= \Big(\int_{|x|>R} |x - P_R(x)|^p d\mu(x)\Big)^{1/p} \\ &= \Big(\int_{|x|>R} (|x| - R)^p d\mu(x)\Big)^{1/p} \leq \epsilon^{1/p} \end{aligned}$$

and by same principles, $d_{W^p}(\mu, (P_R)_{\#}\mu) \leq \epsilon^{1/p}$.

Meanwhile, for any $\varphi \in C_b^0(\mathbb{R}^d)$, we have

$$\begin{aligned} \int \varphi d((P_R)_{\#}\mu_m) &= \int \varphi(P_R(x)) d\mu_m(x) \\ &\to \int \varphi(P_R(x)) d\mu(x) \quad \text{by weak convergence of } (\mu_m) \\ &= \int \varphi d((P_R)_{\#}\mu_m) \end{aligned}$$

So $(P_R)_{\#}\mu_m \xrightarrow{w^*} (P_R)_{\#}\mu$. Then by **Theorem 7.6**, and $\overline{B(0,R)}$ is compact, has $d_{W^p}((P_R)_{\#}\mu_m, (P_R)_{\#}\mu) \to 0$.

Putting these together,

$$\begin{aligned} \limsup_{m\to\infty} d_{W^p}(\mu_m, \mu) &\leq \limsup_{m\to\infty} \Big(d_{W^p}(\mu_m, (P_R)_{\#}\mu_m) + d_{W^p}((P_R)_{\#}\mu_m, (P_R)_{\#}\mu)\Big) + d_{W^p}((P_R)_{\#}\mu, \mu) \\ &\leq 2\epsilon^{1/p} \end{aligned}$$

Let $\epsilon \to 0$ then $d_{W^p}(\mu_m, \mu) \to 0$ as required.

(End of proof) □

7.3 Geodesics in the Wasserstien Space

Definition 7.8) Let $p \in [1, +\infty]$ and (Z, d) be a metric space and $\omega : (a,b)(\subset \mathbb{R}) \to Z$ a curve in Z. We say $\omega \in \mathrm{AC}^p((a,b), Z)$ if $\exists g \in L^p((a,b))$ such that

$$d(\omega(t_0), \omega(t_1)) \leq \int_{t_0}^{t_1} g(s) ds, \quad \forall a < t_0 < t_1 < b$$

If $p = 1$, we say ω is a **absolutely continuous** curve.

If we only have $g \in L^p_{loc}((a,b))$, then we say $w \in \mathrm{AC}^p_{loc}((a,b), Z)$ and curves $\omega \in \mathrm{AC}^1_{loc}((a,b), Z)$ are called **locally absolutely continuous**.

Definition 7.9)

(1) Let (Z,d) be a metric space and $\omega:[0,1]\to Z$ a curve. We define the **length** of ω by

$$\mathrm{Len}(\omega) := \sup\Big\{\sum_{k=0}^{n-1} d(\omega(t_k),\omega(t_{k+1})) : n\in\mathbb{N}, 0=t_0<t_1<\cdots<t_n=1\Big\}$$

(2) A curve $\omega:[0,1]\to Z$ is said to be a **geodesic** between $z_0\in Z$ and $z_1\in Z$ if

$$\omega\in \mathrm{argmin}\Big\{\mathrm{Len}(\overline{\omega})\,\Big|\,\overline{\omega}:[0,1]\to Z, \overline{\omega}=z_0, \overline{\omega}(1)=z_1\Big\}.$$

(3) A curve $\omega:[0,1]\to z$ is said to be a **constant speed geodesic** between $z_0\in z$ and $z_1\in z$ if

$$d(\omega(t),\omega(s)) = |t-s|d(\omega(0),\omega(1))$$

Notes :

(1) If ω is a constant speed geodesic, then it is a geodesic.

(2) If $d(\omega(t),\omega(s)) = |t-s|d(z_0,z_1)$ for all $s,t\in(0,1)$ then $\omega\in \mathrm{AC}^1((0,1),Z)$ with $g(s)=d(z_0,z_1)$.

Definition 7.10) Let (Z,d) be a metric space.

(1) We say (Z,d) is a **length space** if

$$\forall x,y\in z,\quad d(x,y)=\inf\{\mathrm{Len}(\omega):\omega\in \mathrm{AC}^1((0,1),Z),\omega(0)=x,\omega(1)=y\}$$

(2) We say (Z,d) is a **geodesic space** if

$$\forall x,y\in z,\quad d(x,y)=\min\{\mathrm{Len}(\omega):\omega\in \mathrm{AC}^1((0,1),Z),\omega(0)=x,\omega(1)=y\}$$

In particular, the minimum is achieved.

Theorem 7.11) Let $p\in[1,+\infty)$, $X\subset\mathbb{R}^d$ convex. Define $P_t : X\times X\to X$ by $P_t(x,y)=(1-t)x+ty$. Let $\mu,\nu\in\mathcal{P}_p(X)$ and assume that $\pi\in\Pi(\mu,\nu)$ minimize $\mathbb{K}$ with cost $c(x,y)=|x-y|^p$. Then, the curve $\mu_t=(P_t)_\#\pi$ is a constant speed geodesic in $(\mathcal{P}_p(X), d_{W^p})$ connecting μ and ν.

Furthermore, if $\pi=(id\times T)_\#\mu$ where $T:X\to X$ (so in particular T is an optimal transport map), then $\mu_t=((1-t)id+tT)_\#\mu$.

proof) Note $P_0=P^X$ and $P_1=P^Y$, so $\mu_0=(P_0)_\#\pi=(P^X)_\#\pi=\mu$ and similarly $\mu_1=\nu$, so μ_+ connects μ_0 and μ_1. It is enough to show

$$d_{W^p}(\mu_s,\mu_t)=|t-s|d_{W^p}(\mu,\nu)\quad \forall s,t\in[0,1]$$

Suppose $d_{W^p}(\mu_s,\mu_t)\le|t-s|d_{W^p}(\mu,\nu)$ for all $s,t\in[0,1]$. If we can find s,t such that $0\le s<t\le 1$ such that $d_{W^p}(\mu_s,\mu_t)<(t-s)d_{W^p}(\mu,\nu)$ then

$$\begin{aligned} d_{W^p}(\mu,\nu) &\le d_{W^p}(\mu,\nu_s)+d_{W^p}(\mu_s,\mu_t)+d_{W^p}(\mu_t,\nu)\\ &< d_{W^p}(\mu,\nu), \end{aligned}$$

a contradiction. So there are no such s,t.

So we are just left to show that $d_{W^p}(\mu_s,\mu_t) \leq |t-s| d_{W^p}(\mu,\nu)$ for all $s,t \in [0,1]$. To see this, let $\pi_{s,t} = (P_s \times P_t)_\# \pi$. Then for $A \subset X$,

$$\begin{aligned}
\pi_{s,t}(A \times X) &= \pi\Big(\Big\{(x,y) : (P_s \times P_t)(x,y) \in A \times X\Big\}\Big) \\
&= \pi\Big(\Big\{(x,y) : P_s(x,y) \in A\Big\}\Big) \\
&= (P_s)_\# \pi(A)
\end{aligned}$$

Similarly, has $\pi_{s,t}(X \times B) = \mu_t(B)$. Hence $\pi_{s,t} \in \Pi(\mu_s,\mu_t)$. Now

$$\begin{aligned}
d_{W^p}(\mu_s,\mu_t) &\leq \Big(\int_{X\times X} |x-y|^p d\pi_{s,t}(x,y)\Big)^{1/p} \\
&= \Big(\int_{X\times X} |P_s(x,y) - P_t(x,y)|^p d\pi(x,y)\Big)^{1/p} \\
&= |t-s| \Big(\int_{X\times X} |x-y|^p d\pi(x,y)\Big)^{1/p} \\
&= |t-s| d_{W^p}(\mu,\nu)
\end{aligned}$$

The furthermore part of theorem follows from

$$\begin{aligned}
\mu_t &= (P_t)_\# \pi = (P_t)_\# (id \times T)_\# \mu \\
&= (P_t \circ (id \times T))_\mu \mu
\end{aligned}$$

but $P_t \circ (id \times T)(x) = (1-t)x + tT(x)$ so

$$\mu_t = ((1-t)id + tT)_\# \mu$$

(End of proof) □

(8th March, Friday)

8 Gradient Flows in Wasserstein Spaces

We first study the gradient Flow of function defined on an Euclidean space. The theory developed in this lecture can be generalized to the setting in a measure space with a metric.

In Euclidean space, the gradient flow is given by

$$\frac{d}{dt} u(t) = -\nabla \phi(u(t))$$

for $u : t \mapsto \mathbb{R}^d$.

8.1 Gradient Flows for Convex Functions in $\mathbb{R}^d$

Recall, we had

- $\phi : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for all $u_0, u_1 \in \mathbb{R}^d$, $\theta \in [0,1]$

$$\phi(\theta u_0 + (1-\theta)u_1) \leq \theta\phi(u_0) + (1-\theta)\phi(u_1)$$

- If ϕ is differentiable, then it is convex *iff* it has monotonicity of $\nabla\phi$, i.e.

$$\langle \nabla\phi(u_0) - \nabla\phi(u_1), u_0 - u_1 \rangle \geq 0$$

- If $\nabla\phi$ is **L-Lipschitz** then

$$\phi(u_0) \leq \phi(u_1) + \langle \nabla\phi(u_1), u_0 - u_1 \rangle + \frac{L}{2}\left\|u_0 - u_1\right\|^2$$

Definition) ϕ is **λ-convex** if it satisfies, for $u_\theta = (1-\theta)u_0 + \theta u_1$, $\theta \in [0,1]$,

$$\phi(u_\theta) \leq (1-\theta)\phi(u_0) + \theta\phi(u_1) - \frac{\lambda}{2}\theta(1-\theta)\left\|u_0 - u_1\right\|^2$$

If $\phi \in C^2(\mathbb{R}^d)$, this is equivalent to having

(1) $\nabla^2\phi(u) \geq \lambda \mathrm{Id}$. **(Hessian inequality / curvature condition)**

(2) $\langle \nabla\phi(u_0) - \nabla\phi(u_1), u_0 - u_1 \rangle \geq \lambda\|u_0 - u_1\|^2$. **($\lambda$-monotonicity of $\nabla\phi$)**

(3) $\phi(u_1) \geq \phi(u_0) + \langle \nabla\phi(u_0), u_1 - u_0 \rangle + \frac{\lambda}{2}\|u_1 - u_0\|^2$. **(subgradient inequality)**

Remark :

- ϕ is λ-convex *iff* $f(u) := \phi(u) - \frac{\lambda}{2}\|u\|^2$ is convex. Hence whenever ϕ is λ-convex, there are constants a, b such that

$$\phi(x) \geq a + b \cdot x + \frac{\lambda}{2}\|x\|^2$$

- It is direct from the definition of being λ-convexity that whenever ϕ is λ-convex and has a minimizer, then the minimizer is unique.

Consider in the *discrete case*, the problem

$$\min_{u \in \mathbb{R}^d} \phi(u)$$

(I could not copy down) then u should solve $u_{k+1} = u_k - \tau\nabla\phi(u_k)$.

Now let us transfer this to the continuous case. So

$$u_{k+1} = u_k - \tau\nabla\phi(u_k) \quad \Leftrightarrow \quad \frac{u_{k+1} - u_k}{\tau} = -\nabla\phi(u_k)$$

so if we put $u_{k+1} = u_{(k+1)\tau}$, then the analogous equation in the continuous setting will be

$$\frac{d}{dt}u(t) = -\nabla\phi(u(t))$$

Definition) A **Gradient Flow** of $\phi : \mathbb{R}^d \to \mathbb{R}$ starting from an initial pint $u_0 \in \mathbb{R}^d$ is a curve $u : (0, +\infty) \to \mathbb{R}^d$, $t \mapsto u(t) \in \mathbb{R}^d$ that solves (uniquely) the Cauchy problem

$$\begin{cases} \frac{d}{dt}u(t) = -\nabla\phi(u(t)) \\ \lim_{t\to 0^+} u(t) = u_0 \end{cases} \qquad \cdots\cdots\cdots \ (\mathrm{P_{GI}})$$

Proposition 1) Suppose ϕ is a convex function. Let u_1, u_2 be two solutions of $(\mathrm{P_{GI}})$, then we have

$$\frac{d}{dt}\phi(u(t)) = -\left\|\nabla\phi(u(t)))\right\|^2$$

and

$$\|u_1(t) - u_2(t)\| \leq \|u_1(0) - u_2(0)\|$$

i.e. gradient flow is a contraction.

In particular, $(\mathrm{P_{GI}})$ has a unique solution with initial condition given.

proof) The first equality is a direct consequence of $(\mathrm{P_{GI}})$.

To see the second inequality, let $g(t) = \frac{1}{2}\|u_1(t) - u_2(t)\|^2$. Then since $u_j(t) = -\nabla\phi(u_j(t))$ $(j = 1, 2)$,

$$\begin{aligned} g'(t) &= \langle u_1(t) - u_2(t), u_1'(t) - u_2'(t)\rangle \\ &= -\langle u_1(t) - u_2(t), \nabla\phi(u_1(t)) - \nabla\phi(u_2(t))\rangle \leq 0 \end{aligned}$$

by convexity of ϕ. So $g(t) : [0, +\infty) \to \mathbb{R}_+$ has $g(t) \leq g(0)$.

(End of proof) □

We can draw a stronger result if ϕ is λ-convex. Suppose ϕ is λ-convex. Then

$$-\langle u_1(t) - u_2(t), \nabla\phi(u_1(t)) - \nabla\phi(u_2(t))\rangle \leq -\lambda\|u_1(t) - u_2(t)\|^2$$

so

$$g'(t) \leq -\lambda\|u_1(t) - u_2(t)\|^2 = -2\lambda g(t)$$

The *Gronwall's lemma* now implies

$$g(t) \leq e^{-2\lambda t}g(0)$$

and in other words,

$$\|u_1(t) - u_2(t)\|^2 \leq e^{-2\lambda t}\|u_1(0) - u_2(0)\|^2,$$

the *exponential convergence* (also called linear convergence).

Moreover, we have that if ϕ is λ-convex, then $\mathrm{argmin}_u\phi$ is a singleton. So if $u_1(t)$ is a curve and $u_2(t) \equiv \overline{u} = \mathrm{argmin}(\phi)$, then

$$\left\|u_1(t) - \overline{u}\right\|^2 \leq e^{-2\lambda t}\left\|u_1(0) - \overline{u}\right\|^2 \to 0$$

Definition) An **explicit Euler scheme** solves for

$$\frac{u_{k+1}^\tau - u_k^\tau}{\tau} = -\nabla\phi(u_k^\tau)$$

This is very easy to implement. However, if we choose $\tau(1-\frac{\tau L}{2}) > 0$, we get a stability issue.

An **implicit Euler scheme** solves for

$$\frac{u_{k+1}^\tau - u_k^\tau}{\tau} = -\nabla\phi(u_{k+1}^\tau)$$

This corresponds to

$$u_{k+1}^\tau = \text{argmin}\Big\{\phi(u) + \frac{1}{2\tau}\|u - u_k^\tau\|^2\Big\} \quad \Leftrightarrow \quad 0 = \nabla\phi(u_{k+1}^\tau) + \frac{1}{\tau}(u_{k+1}^\tau - u_k^\tau)$$

With such choice of u_k^τ starting from $u_0^\tau = u_0$,

$$\phi(u_{k+1}^\tau) + \frac{1}{2\tau}\Big\|u_{k+1}^\tau - u_k^\tau\Big\|^2 \leq \phi(u_k^\tau)$$

so if we let $\Delta_k = \frac{1}{2\tau}\Big\|u_{k+1}^\tau - u_k^\tau\Big\|^2$, then $\Delta_k \leq \phi(u_k^\tau) - \phi(u_{k+1}^\tau)$, and so

$$\sum_{k+1}^{K} \Delta_k \leq \phi(u_0^\tau) - \phi(u_{K+1}^\tau) =: C < \infty$$

If we have defined

$$\begin{aligned} &u^\tau(t) = u_k^\tau \quad \text{for } t = k\tau \\ &\tilde{u}^\tau(t) = u_k^\tau + (t - k\tau)\nu_{k+1}^\tau \quad \text{for } t \in [k\tau, (k+1)\tau) \text{ and } v_{k+1}^\tau \qquad = \frac{u_{k+1}^\tau - u_k^\tau}{\tau} \end{aligned}$$

then $(\tilde{u}^\tau)'(t) = v^\tau(t)$ and

$$\frac{\|u_{k+1}^\tau - u_k^\tau\|^2}{2\tau} = \frac{\tau}{2}\Big(\frac{\|u_{k+1}^\tau - u_k^\tau\|^2}{\tau^2}\Big)^2 = \frac{\tau}{2}\Big\|v_{k+1}^\tau\Big\|^2 = \frac{\tau}{2}\int_{k\tau}^{(k+1)\tau} \frac{1}{2}\Big\|(\tilde{u}^\tau)'(t)\Big\|^2 dt$$

and because of summability ($\sum_{k+1}^K \Delta_k = C < \infty$),

$$\frac{\tau}{2}\int_0^T \frac{1}{2}\Big\|(\tilde{u}^\tau)'(t)\Big\|^2 < C$$

Using this, one can show that $u^\tau(t)$ converges uniformly to u on every compact set $[0, T]$.

(11th March, Monday)

In Euclidean spaces, we defined the solution u of

$$\frac{du}{dt}(t) = -\nabla\phi(u(t)) \quad \cdots\cdots\cdots \ (*)$$

to be the gradient flow of ϕ. Note,

(a)

$$\frac{d}{dt}\phi(u(t)) = \langle \nabla\phi(u(t)), \frac{du}{dt}(t)\rangle = -\left\|\nabla\phi(u(t))\right\|^2$$

If ϕ is λ-convex, then

(b) $(*)$ is equivalent in $\mathbb{R}^d$ to **Proposition 8.8**

$$\frac{d}{dt}\phi(u(t)) = -\frac{1}{2}\left|\frac{du}{dt}(t)\right|^2 - \frac{1}{2}|\nabla\phi(u(t))|^2 \quad \forall t \in (0,\infty) \quad \cdots\cdots\cdots \text{(EDE)}$$

(c) $(*)$ is equivalent in $\mathbb{R}^d$ to

$$\frac{d}{dt}\left(\frac{1}{2}|u(t)-v|^2\right) + \frac{\lambda}{2}|u(t)-v|^2 \le \phi(v) - \phi(u(t)) \quad \forall t \in (0,\infty),\ \forall v \in \mathbb{R}^d \quad \cdots\cdots\cdots \text{(EVI)}$$

8.2 Gradient Flows in Metric Spaces

Recall : A curve $w : (a,b) \to Z$ is **absolutely continuous** in a metric space (Z,d) if $\exists g \in L^1((a,b))$ such that

$$d(w(t_0), w(t_1)) \le \int_{t_0}^{t_1} g(s)ds \quad \forall t_0, t_1 \in (a,b) \text{ with } t_0 < t_1 \quad \cdots\cdots\cdots (**)$$

And if $g \in L^p((a,b))$ we write $w \in \mathrm{AC}^p((a,b), Z)$.

Definition) We define the **metric derivative** of w by

$$|w'|(t) = \lim_{s\to t} \frac{d(w(s), w(t))}{|t-s|} \quad \cdots\cdots\cdots (\oplus)$$

if the limit exists.

Theorem 8.10) Let (Z,d) be a complete and separable metric space. If $w : (a,b) \to Z$ is absolutely continuous then the limit $(\oplus)$ exists for Lebesgue-a.e. $t \in (a,b)$. Moreover, the function

$$(a,b) \to \mathbb{R}, \quad t \mapsto |w'|(t)$$

is an $L^1(a,b)$-function and one can choose $g = |w'|$ in $(**)$.

If $\tilde{g}$ is any other function satisfying $(**)$, then $|w'|(t) \le \tilde{g}(t)$ for Lebesgue-a.e. t.

Definition 8.11) Let (Z,d) be a metric space then the **metric slope** of $\phi : Z \to \mathbb{R}$ at $v \in \mathbb{R}$ is defined by

$$|\partial\phi|(v) = \limsup_{w\to v} \frac{(\phi(v)-\phi(w))_+}{d(v,w)}$$

8.2.1 GVI Gradient Flows

Definition 8.12) Given a metric space (Z, d) and a function $\phi : Z \to \mathbb{R}$ an **evolution variational inequality (EVI$_\lambda$) gradient flow** is a locally absolutely continuous curve

$$(a, b) \to \mathbb{R}, \quad t \mapsto u(t) \in \mathrm{Dom}(\phi) \subset Z$$

satisfying

$$\frac{1}{2}\frac{d}{dt}\Big(d^2(u(t), v)\Big) + \frac{\lambda}{2} d^2(u(t), v) \le \phi(v) - \phi(u(t)) \quad \text{for a.e. } t \in (0, \infty)$$

Proposition 8.13) Let (Z, d) be a complete and separable metric space and $\phi : Z \to \mathbb{R} \cup \{+\infty\}$ a proper function. If u and v are EVI$_\lambda$ gradient flows with initial condition $u(0) = u_0$ and $v(0) = v_0$. Then

$$d(u(t), v(t)) \le e^{-\lambda t} d(u_0, v_0)$$

Remark : If $\lambda > 0$ then $\exists$ at most one EVI$_\lambda$ gradient flow and $u(t)$ converges exponentially to u^*, the minimiser of ϕ, *i.e.*

$$d(u(t), u^*) \le e^{-\lambda t} d(u_0, u^*)$$

We have not mentioned that ϕ is λ-convex, but if the EVI$_\lambda$ gradient flow exists for any initial condition, then ϕ is λ-convex, and this implies that the minimiser is unique.

(13th March, Wednesday)

Proposition 8.13) Let (Z, d) be a complete and separable metric space and $\phi : Z \to \mathbb{R} \cup \{+\infty\}$ a proper function. If u and v are EVI$_\lambda$ gradient flows with initial condition $u(0) = u_0$ and $v(0) = v_0$. Then

$$d(u(t), v(t)) \le e^{-\lambda t} d(u_0, v_0)$$

proof) We have, for any $w \in \mathbb{R}$,

$$\begin{aligned} \frac{1}{2}\frac{dt}{d} d^2(u(t), w) + \frac{\lambda}{2} d^2(u(t), w) &\le \phi(w) - \phi(u(t)) \quad \cdots\cdots\cdots \ (1) \\ \frac{1}{2}\frac{dt}{d} d^2(v(t), w) + \frac{\lambda}{2} d^2(v(t), w) &\le \phi(w) - \phi(v(t)) \quad \cdots\cdots\cdots \ (2) \end{aligned}$$

Choose $w = v(s)$ in (1) and $w = u(s)$ in (2), then (1), (2) and λ-convexity of u together gives

$$\frac{1}{2}\frac{d}{dt} d^2(u(t), v(s))\Big|_{s=t} + \frac{1}{2}\frac{d}{dt} d^2(v(t), u(s))\Big|_{s=t} \le -\lambda d^2(u(t), v(t))$$

Since

$$\frac{d}{dt} d^2(u(t), v(t)) = \frac{d}{dt} d^2(u(t), v(s))\Big|_{s=t} + \frac{d}{dt} d^2(u(s), v(t))\Big|_{s=t}$$

we have

$$\frac{1}{2}\frac{d}{dt}d^2(u(t),v(t)) \le -\lambda d^2(u(t),v(t))$$

Multiply by $e^{2\lambda t}$, then

$$\frac{d}{dt}\Big(e^{2\lambda t}d^2(u(t),v(t))\Big) = 2\lambda e^{2\lambda t}d^2(u(t),v(t)) + e^{2\lambda t}\frac{d}{dt}d^2(u(t),v(t)) \le 0$$

Hence we have the result.

(End of proof) □

(skipping some materials in the section.)

8.3 Gradient Flows in the Wasserstein Space

We will need the continuity equation (conservation of mass). Assume we have density $\rho(x,t)$ at times t. So each set $A \subset \mathbb{R}^d$ has mass $\int_A \rho(x,t)dx$. We assume that mass is only lost on the boundary of A. So

$$\int_A \frac{\partial \rho}{\partial t}(x,t)dt = \frac{d}{dt}\int \rho(x,t)dx = \int_{\partial A} v(x,t)\cdot n(x,t)dS$$

where the mass is moving with velocity v and n is the unit normal to ∂A. By the divergence theorem,

$$\int_A \frac{\partial \rho}{\partial t}(x,t)dx = -\int_A \nabla\cdot(v(x,t),\rho(x,t))dx$$

This is true for any suitable A, so we have

$$\frac{\partial \rho}{\partial t}(x,t) = -\nabla\cdot(v(x,t)\rho(x,t)),$$

called the **Continuity equation**.

We now make change of notation, $u(t) \to \mu_t$.

8.3.1 Wasserstein Tangent Space

(see [4] Ambrosio, Gigli, Savaré, 2008, “Gradient Flows in Metric Spaces” for more information)

Theorem 8.23) *(Absolutely continuous curves / the continuity equation)*

(1) Let $(0,+\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ be absolutely continuous and $|\mu'| \in L^1((0,+\infty))$ be its metric derivative. Then $\exists$ a vector field $v_t \in L^2(\mu_t,\mathbb{R}^d)$ such that

$$\|v_t\|_{L^2(\mu_t;\mathbb{R}^d)} \le |\mu'|(t) \quad \text{for a.e. } t \in (0,+\infty)$$

and

$$\frac{\partial}{\partial t}\mu_t + \nabla\cdot(v_t\mu_t) = 0 \quad \text{in } \mathbb{R}^d \times (0,+\infty) \quad \cdots\cdots\cdots \ (*)$$

holds in the sense of distributions.

(2) Let $(0,+\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ be continuous with respect to weak* topology on $\mathcal{P}(\mathbb{R}^d)$ and satisfies $(*)$ for some vector field v_t with

$$\int_0^\infty \|v_t\|_{L^2(\mu,\mathbb{R}^d)} dt < +\infty$$

Then μ_t is absolutely continuous and $|\mu'(t)| \leq \|v_t\|_{L^2(\mu_t;\mathbb{R}^d)}$ for a.e. $t \in (0,\infty)$

[We define $(*)$ to hold in the sense of distributions iff

$$\int_0^\infty \int_{\mathbb{R}^d} \Big(\frac{\partial f}{\partial t}(x,t) + v_t(x) \cdot \nabla f(x,t)\Big) d\mu_(x) dt = 0 \quad \forall f \in C_c^\infty(\mathbb{R}^d \times (0,+\infty))$$

]

The following proposition motivates the tangent space. The tangent direction can be thought of as the direction in which a variation in the direction stays in the probability space.

Proposition 8.26) Let $(0,+\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ be absolutely continuous. Suppose $v_t \in \overline{\{\nabla\varphi : \varphi \in C_c^\infty(\mathbb{R})\}}^{L^2(\mu_t;\mathbb{R}^d)}$ satisfies $(*)$. Assume there is an optimal transport map $T^{(t,s)}$ between μ_t and ν_s for each s,t, *i.e.* $T^{(t,s)}_\# \mu_t = \mu_s$. Then

$$\lim_{h\to 0} \frac{T^{(t,t_h)} - id}{h} = v_t$$

where this limit is taken in $L^2(\mu_t;\mathbb{R}^d)$ (for all t? check)

Definition 8.27) For $\mu \in \mathcal{P}_2(\mathbb{R}^2)$, the **tangent space to $\mathcal{P}_2(\mathbb{R}^d)$ at the point μ** is

$$\mathrm{Tan}_\mu \mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\varphi; \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu;\mathbb{R}^d)}$$

We can use **Proposition 8.26** to differentiate $t \mapsto \frac{1}{2} d^2_{W^2}(\mu_t,\sigma)$.

Theorem 8.28) Let $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$, $(0,+\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ be absolutely continuous and $v_t \in \mathrm{Tan}_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$ satisfy $(*)$. Assume there is an optimal transport map T^t between σ and μ_t for all t, *i.e.* $T^{(t)}_\# \mu_t = \sigma$. Then

$$\frac{1}{2}\frac{d}{dt} d^2_{W^2}(\mu_t,\sigma) = \int_{\mathbb{R}^d} (x - T^{(t)}(x)) \cdot v_t(x) d\mu_t(x)$$

sketch proof) Maps $T^{(t)}$, $T^{(t,t+h)}$, $T^{(t+h)}$ are all optimal maps. We suppose $T^{(t)} = T^{(t+h)} \circ T^{(t,t+h)}$. *[Of course, this is not true in general, but there is a reasonable justification for this : when h is small, then composition of optimal maps would be close to an optimal map, so we expect the equality to hold up to a $o(h)$ correction.]* Then

$$\begin{aligned}
\frac{d}{dt} d^2_{W^2}(\mu_t,\sigma) &= \lim_{h\to 0^+} \frac{1}{h}\Big[\int_{\mathbb{R}^d} |T^{(t+h)}(x) - x|^2 d\mu_{t+h}(x) - \int_{\mathbb{R}^d} |T^{(t)}(x) - x|^2 d\mu_t(x)\Big] \\
&= \lim_{h\to 0^+} \frac{1}{h} \int_{\mathbb{R}^d} \Big[|T^{(t+h)}(T^{(t,t+h)}(x)) - T^{(t,t+h)}(x)|^2 - |T^{(t)}(x) - x|^2\Big] d\mu_t(x) \\
&= \lim_{h\to 0^+} \int_{\mathbb{R}^d} (x - T^{(t,t+h)}(x)) \cdot (2T^{(t)}(x) - x - T^{(t,t+h)}(x)) d\mu_t(x) \\
&= \int_{\mathbb{R}^d} (-v_t(x)) \cdot (2T^{(t)}(x) - x - x) d\mu_t(x)
\end{aligned}$$

(End of proof) □

(15th March, Friday)

8.3.2 Gradient Flow in $(\mathcal{P}_2(\mathbb{R}^d), d_{W^2})$ and Evolutionary PDEs

Consider function $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. Take the Gradient flow in the Wasserstien space. We would end up a evolutionary PDE of the form

$$\frac{\partial \mu}{\partial t} - \nabla \cdot (\mu \nabla \frac{\delta \phi}{\delta \mu}) = 0 \quad \cdots\cdots\cdots \ (*)$$

where $\frac{\delta \phi}{\delta \mu}$ is the **1st variation** of ϕ, *i.e.*

$$\langle \frac{\delta \phi}{\delta \mu}, \chi \rangle = \frac{d}{d\epsilon} \phi(\mu + \epsilon \chi)|_{\epsilon=0}, \quad \forall \chi \in \mathcal{P}_2(\mathbb{R}^d)$$

Alternatively,

$$\langle \frac{\delta \phi}{\delta \mu}, \chi \rangle = \lim_{h \to 0^+} \frac{\phi(\mu + h\chi) - \phi(\mu)}{h}$$

It is also not clear how to take the divergence of a measure, but in the cases we care about, we can work with densities.

We will consider the special case in which ϕ adopts the following decomposition :

$$\phi(\mu) = u(\mu) + v(\mu)$$

where

$$\begin{aligned} u(\mu) &= \begin{cases} \int_{\mathbb{R}^d} \overline{u}(\rho(x))dx & \text{if } \mu \ll \mathcal{L}^d, \rho = \frac{d\mu}{d\mathcal{L}^d} \\ +\infty & \text{otherwise} \end{cases} \qquad \cdots\cdots\cdots \ (\oplus) \\ v(\mu) &= \int_{\mathbb{R}^d} \overline{v}(x) d\mu(x) \end{aligned}$$

One can show (an exercise)

$$\frac{\delta \phi}{\delta \mu}(\mu) = \overline{u}'(\rho) + \overline{v}$$

Say $\overline{u}(r) = r \log r$, then $\frac{\delta \phi}{\delta \mu}(\mu) = \log \rho + 1 + \overline{v}$. In this case, $(*)$ is understood as

$$\frac{\partial \rho}{\partial t} - \nabla \cdot (\rho \nabla (\log \rho + \overline{v} + 1)) = \frac{\partial \rho}{\partial t} - \nabla \cdot (\nabla \rho + \rho \nabla \overline{v}) = 0$$

Theorem 8.30) *(Existence of Gradient Flows)* Let $\overline{v}$ and $\overline{u}$ satisfy

(1) $\overline{v} : \mathbb{R}^d \to \mathbb{R}$ is λ-convex.

(2) $\overline{u} : [0, \infty) \to \mathbb{R}$ convex, $\overline{u}(0) = 0$, $\liminf_{s \to 0^+} \frac{\overline{u}(s)}{s^\alpha} > -\infty$ for some $\alpha > \frac{d}{d+2}$, $\lim_{s \to \infty} \frac{u(s)}{s} = +\infty$, $s \mapsto s^d u(s^d)$ is convex and non-increasing.

Let $\phi = u + v$, where u, v are given by $(\oplus)$. Then for all $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique $\{\mu_t\}_{t \in (0,\infty)} \subset \mathrm{Lip}_{loc}((0, +\infty); \mathcal{P}_2(\mathbb{R}^d))$ (Lipschitz functions w.r.t. Wasserstein metric) and $v_t \in \mathrm{Tan}_{\mu_t}(\mathcal{P}_2(\mathbb{R}^d))$ such that $\lim_{t \to 0^+} \mu_+ = \mu_0$ in $\mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \int_{\mathbb{R}^d} |v_t|^2 d\mu_t = |\mu_t'|^2$ is $L^\infty_{loc}((0, +\infty))$,

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0 \quad \text{in } (0, +\infty) \times \mathbb{R}^d$$

and $\forall \sigma \in \mathrm{Dom}(\phi)$,

$$\frac{1}{2}\frac{d}{dt}d^2_{W^2}(\mu_t,\sigma)+\frac{\lambda}{2}d^2_{W^2}(\mu_t,\sigma)\le \phi(\sigma)-\phi(\mu_t)$$

and

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}\Big(\langle v_t(x),x-y\rangle+\frac{\lambda}{2}|x-y|^2\Big)d\pi_t(x,y)\le\phi(\sigma)-\phi(\mu_t)\quad\cdots\cdots\cdots\ (**)$$

for all $\pi_t\in\Pi_{\mathrm{opt}}(\mu_t,\sigma)$ where $\Pi_{\mathrm{opt}}(\mu_t,\sigma)$ is the set of transport plans in $\Pi(\mu_t,\sigma)$ that is optimal for the Kantorovich problem with quadratic cost.

Not proving this theorem.

Application : Let $\overline{u}(\rho)=\rho\log\rho$. Fix $\zeta\in C_c^\infty(\mathbb{R}^d)$, define $T_\epsilon=id+\epsilon\nabla\zeta$. *[Exercise : if $\epsilon\max_{\mathbb{R}^d}\|D^2\zeta\|<1$ then $\pi_\epsilon=(id\times T_\epsilon)_\#\mu_t$ is an optimal transport plan (with quadratic cost) between μ_t and $\mu_t^{(\epsilon)}=(T_\epsilon)_\#\mu_t$.]* Choose $\sigma=(T_\epsilon)_\#\mu_t$ in $(**)$, then

$$\mathrm{LHS}=\int_{\mathbb{R}^d\times\mathbb{R}^d}\Big(\langle v_t(x),x-y\rangle+\frac{\lambda}{2}|y-x|^2\Big)d(id\times T_\epsilon)_\#\mu_t(x,y)\le\phi(\mu_t^\epsilon)-\phi(\mu_t)=\mathrm{RHS}$$

We have

$$\begin{aligned}\mathrm{LHS}&=\int_{\mathbb{R}^d}\Big(\langle v_t(x),x-T_\epsilon(x)\rangle+\frac{\lambda}{2}|T_\epsilon(x)-x|^2\Big)d\mu_t(x)\\&=\int_{\mathbb{R}^d}\Big(-\epsilon\langle v_t(x),\nabla\zeta(x)\rangle+\frac{\epsilon^2\lambda}{2}|\nabla\zeta(x)|^2\Big)d\mu_t(x)\\&\ge-\epsilon\int_{\mathbb{R}^d}\langle v_t(x),\nabla\zeta(x)\rangle d\mu_t(x)\end{aligned}$$

Let $\rho_t=\frac{d\mu_t}{d\mathcal{L}^d}$ and $\rho_t^{(\epsilon)}=\frac{d\mu_t^{(\epsilon)}}{d\mathcal{L}^d}$, so

$$\mathrm{RHS}=\int_{\mathbb{R}^d}\Big(\log(\rho_t^{(\epsilon)}(x))+\overline{v}(x)\Big)d\mu_t^{(\epsilon)}(x)-\int_{\mathbb{R}^d}\Big(\log(\rho_t(x))+\overline{v}(x)\Big)d\mu_t(x)$$

By change of variable formula, has

$$\begin{aligned}\rho_t(x)&=\rho_t^{(\epsilon)}(T_\epsilon(x))|\det(\nabla T_\epsilon(x)))|\\&=\rho_t^{(\epsilon)}(T_\epsilon(x))|\det(id+\epsilon D^2\zeta(x))|\end{aligned}$$

So

$$\begin{aligned}\mathrm{RHS}&=\int_{\mathbb{R}^d}\Big(\log(\rho_t^{(\epsilon)}(T_\epsilon(x)))-\log\rho_+(x)\Big)d\mu_t(x)+\int_{\mathbb{R}^d}(\overline{v}(T_\epsilon(x))-\overline{v}(x))d\mu_t(x)\\&=-\int_{\mathbb{R}^d}\log|\det(id+\epsilon D^2\zeta(x))|d\mu_+(x)+\int_{\mathbb{R}^d}\overline{v}(T_\epsilon(x))-\overline{v}(x)d\mu_t(x)\end{aligned}$$

Hence

$$-\int_{\mathbb{R}^d}\langle v_t(x),\nabla\zeta(x)\rangle d\mu_t(x)\le-\frac{1}{\epsilon}\int_{\mathbb{R}^d}\log|\det(id+\epsilon D^2\zeta(x))|d\mu_t(x)+\frac{1}{\epsilon}\int_{\mathbb{R}^d}\overline{v}(T_\epsilon(x))-\overline{v}(x)d\mu_t(x)$$

We have

$$\begin{aligned}\frac{\overline{v}(T_\epsilon(x))-\overline{v}(x)}{\epsilon}&=-\frac{1}{\epsilon}(\nabla\overline{v}(x)\cdot(T_\epsilon(x)-x)+o(\epsilon))\\&=\nabla\overline{v}(x)\cdot\nabla\zeta(x)+o(\epsilon)\\&\to\nabla v(x)\cdot\nabla\zeta(x)\quad\text{as }\epsilon\to0\end{aligned}$$

Let $F_D(\epsilon) = \log|\det(id + \epsilon D)|$. We have (exercise)

$$\frac{dF_D}{d\epsilon} = \operatorname{tr}((id + \epsilon D)^{-1} D)$$

Hence

$$\begin{aligned}
\frac{1}{\epsilon}(F_{D^2\zeta(x)}(\epsilon) - F_{D^2\zeta(x)}(0)) &= \frac{dF_{D^2\zeta(x)}}{d\epsilon}(0) + o(\epsilon) \\
&= \triangle\zeta(x) + o(\epsilon) \\
&\to \triangle\zeta(x) \quad \text{as } \epsilon \to 0
\end{aligned}$$

Now we have

$$-\int_{\mathbb{R}^d} \langle v_t(x), \nabla\zeta(x)\rangle d\mu_t(x) \leq \int_{\mathbb{R}^d} (-\triangle\zeta(x) + \nabla\overline{v}(x) \cdot \nabla\zeta(x)) d\mu_t(x) \quad \forall\zeta \in C_c^\infty(\mathbb{R}^d)$$

Also, by replacing ζ with $-\zeta$, we have the inequality in the converse direction, so in fact

$$-\int_{\mathbb{R}^d} \langle v_t(x), \nabla\zeta(x)\rangle d\mu_t(x) = \int_{\mathbb{R}^d} (-\triangle\zeta(x) + \nabla\overline{v}(x) \cdot \nabla\zeta(x)) d\mu_t(x) \quad \forall\zeta \in C_c^\infty(\mathbb{R}^d)$$

This is a weak form of $-v_t\mu_t = \nabla\mu_t + \mu_t\nabla\overline{v}$, and together with the fact that $\frac{d}{dt}\mu_t + \nabla\cdot(v_t\mu_t) = 0$, we conclude that

$$\frac{d\mu_t}{dt} - \nabla \cdot (\mu_t \nabla v + \nabla\mu_t) = 0$$

which is in accordance with our previous intuitive result.