Introduction to Optimal Transport

Lectured by Matthew Thorpe

Lent 2019

1 Background on Measure Theory

Definition) σ-algebra, measure, probability measure, Borel σ-algebra on topological space X, probability measure, Polish space

Definition 1.1.) weak-* convergence of measures

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) tight probability measures

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) pushforward of measure

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)* $(S \circ T)_\#\mu = S_\#(T_\#\mu)$.

Theorem 1.6) *(Disintegration of Measures)* Let $\mathbb{X}, Z$ be Polish spaces and $P : \mathbb{X} \to Z$ Be measurable. Let $\pi : \mathcal{P}(\mathbb{X})$ 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

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

(not proving)

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

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

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 $\nu \in \mathcal{P}(Y)$ if...

Existence of Transport Map

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

The Monge's Problem

Definition 2.4) Monge's Optimal Transport Problem

2.2 The Kantorovich Formulation

Definition) $\Pi(\mu, \nu)$, Kantorovich Optimal Transport Problem

We have $\inf \mathbb{K}(\pi) \leq \inf \mathbb{M}(T)$ (- prove this)

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

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

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

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

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

(the implication is an exercise) 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,

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

(proof uses proposition 3.3)

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 problem, *i.e.* $T^{\dagger}_{\#}\mu=\nu$ and $\inf_{T:T_{\#}\mu=\nu}\mathbb{M}(T)=\mathbb{M}(T^{\dagger})$.

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

Definition) We say $\Gamma\subset\mathbb{R}^2$ **is monotone with respect to** d if...

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. $\text{supp}(\pi)$ is *cyclically monotone.*

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 $m=n$?

(2) Do all $\{x_i\},\{y_i\}$ need to be distinct?

Definition) Extreme points of a convex set $B\subset M$, M a Banach space

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,

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

(not proved in lecture)

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 not done in lecture)

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.

4 Kantorovich Duality

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

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

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) Legendre-Fenchel transformation of a function on a normed vector space.

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.
(uses lemma 4.4 and theorem 4.5)

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 $\overline{D} = \overline{\text{int}(D)}$

(was an 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)

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

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

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

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

4.4 Existence of Maximisers to the Dual Problem

Definition 4.9) c-transform of $\varphi \to \mathbb{R} \cup \{+\infty\}$.

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 proof uses lemma 4.11 and 4.12)

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 $\mathbb{J}(\varphi, \psi) \geq \mathbb{J}(\tilde{\varphi}, \tilde{\psi})$ and $\varphi(x) + \psi(y) \leq c(x, y)$ 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 $\varphi \leq c_X$ and $\psi \leq c_Y$ then $(\varphi, \psi) \in \Phi_c$.

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 $c(x, y) \leq c_X(x) + c_Y(y)$ 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 $\varphi_k(x) \leq c_X(x)$ for μ-a.e. x and $\psi_k(y) \leq c_Y(y)$ for ν-a.e y for all $k \in \mathbb{N}$.

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 $\|f\|_{Lip} = \sup_{x\neq y\in X} \frac{|f(x)-f(y)|}{c(x,y)}$ 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

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

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...

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.

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

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

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

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

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 $c(x,y) = |x-y|^2$.

6 Existence and Characterisation of Transport Maps

6.1 Knott-Smith Optimality and Beiner's Theorem

Definition) subdifferential of a convex function φ

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

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.*

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

6.2 Preliminary Results for Convex Analysis

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

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 $\forall x, y \in \mathbb{R}^d$,

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

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 uses Rademacher's theorem

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. (not prove this results.)

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

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 $\|\tilde{\varphi}^* - f\|_{L^1(\mu)} = \int_Y \big(\tilde{\varphi}^*(y) - f(y)\big) d\nu(y)$. Hence

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

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 $\tilde{\Phi}$.

6.4 Proof of Brenier's Theorem

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 $\tilde{\Phi}$.

7 Wasserstein Distances

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

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

7.1 Wasserstein Distances

We work on the space of measures with bounded p^{th} moment, $\mathcal{P}_p(X)$.

Definition 7.2) p -Wasserstien distance

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.

-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 $P^{X,Y}(x,y,z) = (x,y)$, $P^{Y,Z}(x,y,z) = (y,z)$.

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 $d_{W^p}^p(\mu,\nu) \leq \operatorname{diam}(X)^{p-1} d_{W^1}(\mu,\nu)$ (where $\operatorname{diam}(X) = \sup_{w,z\in X} |w - z|$)

7.2 The Wasserstein Topology

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

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

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

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

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..., absolutely continuous curve, locally absolutely continuous.

Definition 7.9) length of ω, geodesic between $z_0, z_1 \in Z$, constant speed geodesic

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) (Z, d) a metric space, Define length space if... is a geodesic space if...

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