Optimal Transport and Machine Learning

November 19, 2021

1 Introductory notes

Monge 1781.

Suppose μ and ν are two measures on $\mathbb{R}^d, d \geq 1$.

Consider any function $T : \mathbb{R}^d \to \mathbb{R}^d$ that push forward μ to ν.

Suppose $X \sim \mu$, then $y = T(X) \sim \nu$.

Problem 1.1. Monge's problem. What is the infimum of

$$\int ||T(x) - x||\mu(dx) = \mathbb{E}[||T(X) - X||]$$

over the set of all push forwards of μ to ν?

Monge's idea: move dirt to castle.

$$Vol(Dirt) = Vol(Castle)$$

Every x in Dirt should be carried to y. We wish to have minimum work possible. Two points are $||y - x||$.

Summing up all the things,

$$\inf \int ||T(x) - x||\mu(dx)$$

This is a hard problem.

Consider if we just take $\mu = \delta_0$, and $\nu = Ber(1/2)$.

The set of pushforwards is not nice (not convex, smooth, ...)

How to generalize? Monge is mapping cost as $||T(x) - x||$ = cost of transporting.

Why not use $||T(x) - x||^2$? Why note use $||T(x) - x||^{40}$?

Define a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$

Generalized Monge Problem (MP):

find

$$\inf \int c(x, T(x))\mu(dx)$$

among all pushforwards of μ to ν.

Kantorovich's relaxation: without enforcing existence of mapping.

Coupling Given μ and ν, a coupling of (μ, ν) refers to any joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$,

such that if $(X, y) \sim \rho$, then $X \sim \mu$, $Y \sim \nu$.

Example 1.2. Suppose T is a pushforward from μ to ν, then $(X, T(X))$ where $X \sim \mu$ is a coupling of (μ, ν).

Example 1.3. Suppose $X \sim \mu$ independent of $Y \sim \nu$, then $(X, Y) \sim \mu \otimes \nu$ is a coupling of (μ, ν).

Let $\pi(\mu, \nu)$ be the set of couplings, then $\pi(\mu, \nu) \neq \emptyset$.

Problem 1.4. Kantorovich Problem (KP).

Find

$$\inf_{\pi \in \Pi(\mu,\nu)} \int c(x, y) d\pi$$

E.g.

$$\inf_{\pi \in \Pi(\mu,\nu)} \int ||x - y||^2 \pi(dxdy)$$

Advantage

1. $\pi(\mu, \nu)$ is a non-empty convex set.
2. The function being optimized is affine.
3. KP is a linear programming problem.

In details:

1. $\pi(\mu, \nu)$ is convex. How to verify $\pi \in P(\mathbb{R}^d \times \mathbb{R}^d)$ is an element in $\Pi(\mu, \nu)$?
Take some $A \subseteq \mathbb{R}^d$, sample $(X, Y) \sim \Pi$,
check:

$$P_\Pi(x \in A) = \mu(A), P_\Pi(y \in A) = \nu(A), \forall A.$$

Alternatively, consider f to be a bounded function,

$$c_x^f := \int f(x) d\mu, \quad c_y^f := \int f(y) d\nu$$

$$\bar{f}(x, y) := f(x), \quad \underline{\text{f}}\text{(x,y):=f(y)}$$

Check

$$\begin{cases} \mathbb{E}_\Pi[\bar{f}] = \int \bar{f}(x, y) d\pi = c_x^f \\ \mathbb{E}_\Pi[\underline{\text{f}}] = \int \underline{\text{f}}\text{(x,y)d}\pi = c_y^f \end{cases}$$

Intersecting $P(\mathbb{R}^d \times \mathbb{R}^d)$.

2. The function is linear in π

How are MP and KP related?

What is the value of problem?

Is $\inf = \min$**? Does solution exist?**

Is the minimizer unique?

If so, how does the optimizer look like? Will focus mostly on $c(x, y) = ||y - x||^2$.

1.1 When is the infimum achieved?

Weienstrass Theorem.

Theorem 1.5. *Suppose the cost function c is continuous, then KP admits a solution. That is, there is some coupling* $\pi^* \in \Pi(\mu, \nu)$ *that attains infinum.*

Proof. Depends on this basic lemma. □

Lemma 1.6. *If* f *is a real-valued continuous function on a compact metric space* X*, then* $\exists$ *some* $x^* \in X$ *such that*

$$f(x^*) = \min_{x \in X} f(x)$$

Proof. Let $l = \inf_x f(x)$. Assume $l > -\infty$.

For every $n \geq 1$, $\exists$ some x_n s.t.

$$l \leq f(x_n) \leq l + \frac{1}{n}$$

Then sequence $(x_n, n \geq 1)$ has a converging subsequence.

$$x_{n_k} \to x^*$$

What is $f(x^*)$?

$$f(x^*) = \lim_{k \to \infty} f(x_{n_k}) \leq \lim_{n \to \infty} (l + \frac{1}{n_k}) = l = \inf_x f(x)$$

□

Metrics on probability measures. $P(\mathbb{R}^n)$

Definition 1.7. For a sequence $(\rho_k, k \geq 1)$ in $P(\mathbb{R}^n)$, say $\lim_{k \to \rho_k} = \rho$ if

$$\lim \int f d\rho_k = \int f d\rho, for\ all\ bounded\ continuous\ functions\ f : \mathbb{R}^n \to \mathbb{R}$$

"Weak convergence of probability measures" There is a metric that gives us this weak convergence.

$$d(\rho_0, \rho_1) = \sup_{f \in BL} | \int f d\rho_0 - \int f d\rho_1 |$$

BL is the set of all functions bounded (B) by 1 and is Lipschitz (L). $|f(x)| \leq 1$, $|f(x) - f(y)| \leq ||x - y||$.

Theorem 1.8. *For any μ and ν, the set*
$\Pi(\mu, \nu)$ is compact in the topology of weak convergence.

Proof follows from Prokhonor's Theorem. We can verify from this theorem (for stating out what is weak convergence).

Thus, $\Pi(\mu, \nu)$ is a compact metric space.

The entire $P(\mathbb{R}^n)$ cannot be compact.

$$\rho_k = \delta_k, \lim_{k \to \infty} \int f(x) d\rho_k = f(k)$$

Proof. [Sketch]

Assume μ, ν are compactly supported. It means there exist a big compact ball in $\mathbb{R}^d$ that the entire measures live in this compact ball.

Every element in $\Pi(\mu, \nu)$ must be supported in some big enough box $[-a, a]^{2d}$.

On that box, the continuous cost function c is also bounded.

Thus,

$$\pi \in \Pi(\mu, \nu) \to \int c(x, y) d\pi$$

is a continuous function.

By Weienstrass, $\exists \pi^*$,

$$\inf_{\pi \in \Pi(\mu, \nu)} \int c d\pi = \int c(x, y) d\pi^*.$$

□

1.2 Linear Algebra

Suppose

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

$$\nu = \frac{1}{n} \sum_{j=1}^{n} \delta_{y_j}$$

What is $\Pi(\mu, \nu)$? Given by Doubly-Stochastic matrices (DS matrices).

Definition 1.9. $A_{n\times n} = (a_{ij})$ is DS if

1. $a_{ij} \geq 0$
2. Row sum $= 1$
3. Col sum $= 1$

$\frac{1}{n}A \iff \Pi(\mu,\nu)$.
$P(X = x_i, Y = y_j)$.
Special case: Permutation matrices. 1-2, 2-1, 3-3.
$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$$(\frac{1}{n}A_\pi) \iff Push\ Forwards$$

KP in Linear Algebra.

$$C = (c_{ij}),\ c_{ij} = c(x_i, y_j)$$

$$\frac{1}{n}\langle A, C\rangle = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}c_{ij}$$

KP becomes

$$\inf_{A\ over\ all\ DS\ matrices\ n\times n} \langle A, C\rangle$$

Fact 1.10. *This mimimum exists and is achieved at some permutation matrix.*

$$(KP) = (MP)$$

$$\mu = \sum p_i\delta_{x_i}, \nu = \sum q_j\delta_{y_j}$$

Find $\Pi(\mu,\nu)$ is some set of matrices

$$\inf_A \langle C, A\rangle$$

is a Linear programming problem.

2 Convex functions and their duals

2.1 Review

MK OT problem

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

Given μ, ν on $\mathbb{R}^d$

$$\pi(\mu,\nu) - set\ of\ couplings$$

KP is

$$\inf_{\pi\in\Pi(\mu,\nu)} \int ||y-x||^2 d\pi$$

If this infimum is given by a coupling $(X, T(X)), X \sim \mu, T(X) \sim \nu$. We say KP admits a Monge solution.

Example 2.1. $\mu = \mathcal{N}(0,I)$ on $\mathbb{R}^d$. $\nu = \mathcal{N}(w,I)$ on $\mathbb{R}^d$. What is the solution of KP?

The solution is a shift that

$$T(x) = x + w$$

Here, $(Z, T(Z))$ is the optimal solution to (KP).

How do I argue this? Brenier Theorem.

The reason is $T(X) = \nabla f(x), f(x) = \frac{1}{2}||x+w||^2$. If you can find a convex function gradient, this must be the optimal.

If μ has a density (absolutely continuous), no matter what ν is, there always exists some convex function f, ∇f pushforwards μ to ν.

Weak convergence of measures $(\rho_k, k \geq 1)$ seq. in $P(\mathbb{R}^d)$

Say $\rho_k \to \rho$ if

$$\int f d\rho_k = \int f d\rho$$

For every bounded continuous function $f : \mathbb{R}^n \to \mathbb{R}$.

Example 2.2. From $[0,1]$, draw k partitions.

$$\rho_k = Unif[\frac{i}{k}, i = 1, 2, ..., k]$$

When $k \to \infty$,

$$\rho_k \to \rho = Unif[0,1]$$

Why is this true? Take any f bounded and continuous.

$$\int f d\rho_k = \sum f(i/k)\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} f(i/k) =_{k\to\infty} \int_0^1 f(x)dx = \int f(x)\rho(dx)$$

Even $X_1, ..., X_k \sim_{iid} Unif[0,1]$.

$$\frac{1}{k}\sum_{i=1}^{k}\delta_{X_i} \to_{a.s.}^{k\to\infty} Unif[0,1]$$

2.2 Convex Analysis

Definition 2.3. $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called convex if for any $x, y \in \mathbb{R}^d$, any $0 < t < 1$

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

And strictly convex if

$$f((1-t)x + ty) < (1-t)f(x) + tf(y)$$

Definition 2.4. f is concave if $-f$ is convex.

Definition 2.5. A is a convex set, if $x, y \in A$, then

$$\{(1-t)x + ty, 0 \leq t \leq 1\} \subseteq A.$$

Example 2.6. $x \in \mathbb{R}^d$, $f(x) = ||x||^2$ strictly convex.

Example 2.7. If $f(x) = \sum_i |x_i|$. This is convex but not strictly convex.

Example 2.8. $f(x) = ||x||_p^p, p > 1$, is strictly convex. If $p < 1$, concave function.

Example 2.9. $f(x) = \log\left(\sum_{i=1}^d e^{X_i}\right), x \in \mathbb{R}^d$.
Verify this is convex. Show the Hessian.

Convex functions could be infinity somewhere

Example 2.10. $f(x) = \begin{cases} -\log x & x > 0 \\ +\infty & x \leq 0 \end{cases}$
This is also a convex function.

Domain of $f = \{x \in \mathbb{R}^d : f(x) < +\infty\} \neq \emptyset$.

2.2.1 How convex sets related to convex function

Suppose Ω is a convex set.

Convex indicator function: $f(x) = \begin{cases} 0, & x \in \Omega \\ +\infty, & x \notin \Omega \end{cases}$
Verify that f is convex function if Ω is convex set.
Conversely, convex functions to convex sets.
Suppose f is a Convex function. Consider the epigraph of f

$$epi(f) = \Omega = \left\{(x,t) \in \mathbb{R}^{d+1} : t \geq f(x)\right\}$$

f is convex function if and only if the epigraph is convex set.

Properties

1. Closed under supremum.

$$\{f_\alpha, \alpha \in I\}$$

$$f_\alpha \to \mathbb{R} \cup \{\infty\}$$

is convex, then so is

$$f(x) = \sup_\alpha f_\alpha(x).$$

$x, y, 0 < t < 1$

$$f_\alpha((1-t)x + ty) \leq (1-t)f_\alpha(x) + tf_\alpha(y)$$

Then

$$\sup_\alpha f_\alpha((1-t)x + ty) \leq \sup_\alpha [(1-t)f_\alpha(x) + tf_\alpha(y)]$$

2. Convex functions may not be always differentiable, or continuous.
$f(x) = \begin{cases} x^2, & -1 < x < 1 \\ 2, & x = \pm 1 \\ \infty, & |x| > 1 \end{cases}$
This function is convex but not continuous at the boundary.
It is locally Lipschitz in the $interior(dom(f))$
It is differentiable almost everywhere inside $interior(dom(f))$.
It is "double differentiable" a.s.

We are only going to consider a convex function that are lower semicontinuous

$$(x_k) \to x$$
$$\lim f(x_k) \geq f(x) \iff epi(f) \; is \; closed.$$

3. Every convex lower semicontinuous function can be written in the following representation

$$\exists (a_\alpha \in \mathbb{R}^d, b_\alpha \in \mathbb{R}, \alpha \in I)$$

such that

$$f(x) = \sup_\alpha \left[\underbrace{\langle a_\alpha, x \rangle + b_\alpha}_{affine \; in \; x} \right]$$

This is a dual representation of f.

Definition 2.11. Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$. Define Legendre transform (convex conjugate) of f,

$$f^*(y) = \sup_{x \in \mathbb{R}^d} [\langle x, y \rangle - f(x)] \to convex \; l.s.c \; function$$

Double star?

$$f^{**}(x) = \sup_{y \in \mathbb{R}^d} [\langle x, y \rangle - f^*(y)]$$

$$f^{**} = f \iff f \; is \; convex + lsc$$

Otherwise, f^{**} is called "convex evelope".

Example 2.12. $f(x) = \frac{1}{2}||x||^2$.

$$f^*(x) = \sup_x \left[\langle x, y \rangle - \frac{1}{2}||x||^2 \right]$$

Let $g(x) = \langle x, y \rangle - \frac{1}{2}||x||^2$

$$\nabla g(x) = y - x = 0$$

Therefore,

$$f^*(x) = \frac{1}{2}||y||^2 = f(y)$$

$f = f^*$ is self-dual.

Example 2.13. $f(x) = \begin{cases} -\log x, & x > 0 \\ +\infty, & x \leq 0 \end{cases}$

$$f^*(y) = \begin{cases} -1 - \log |y|, & y < 0 \\ +\infty, & y \geq 0 \end{cases}$$

What if we have a f^{**}? Since $f(x)$ is convex and lsc, we get back $f(x)$.

Example 2.14. $\Omega = [-1,1]^d$, $f(x) = \begin{cases} 0, & x \in \Omega \\ +\infty, & x \notin \Omega \end{cases}$

$$\begin{aligned} f^*(y) &= \sup_{x \in \mathbb{R}^d} [\langle x, y \rangle - f(x)] \\ &= \sup_{x \in \Omega} [\langle x, y \rangle] \\ &= \sup_{x \in [-1,1]^d} \sum_{i=1}^{d} x_i y_i \\ &= ||y||_1 \end{aligned}$$

$$\begin{aligned} f^{**}(x) &= \sup_{y \in \mathbb{R}^d} [\langle x, y \rangle - f^*(y)] \\ &= \sup_{y \in \mathbb{R}^d} [\langle x, y \rangle - ||y||_1] \\ &= \begin{cases} +\infty, & if\ x \notin \Omega \\ 0, & if\ x \in \Omega \end{cases} \end{aligned}$$

Interestingly, if $\Omega = (-1,1)^d$, $f^{**}(x) = [-1,1]^d$.

Theorem 2.15. *Suppose f and f^* are convex and differentiable over $\mathbb{R}^d$. (Differentiable implies lsc).*

1. $f(x) + f^*(y) \geq \langle x, y \rangle$ for all $x, y \in \mathbb{R}^d$, with $=$ holds if and only if $y = \nabla f(x)$.
2. $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$, $\nabla f^* : \mathbb{R}^d \to \mathbb{R}^d$ are inverse of one another.

$$\nabla f(\nabla f^*(y)) = y$$

$$\nabla f^*(\nabla f(x)) = x$$

Proof. Idea of the proof.
(1)

$$f^*(y) = \sup_x \left[\langle x, y\rangle - f(x)\right] \geq \langle x, y\rangle - f(x)$$

$$f(x) + f^*(y) \geq \langle x, y\rangle$$

Where the supremum is achieved?
FO condition:

$$y = \nabla f(x)$$

$$f^*(y) = \langle x, y\rangle - f(x), y = \nabla f(x).$$

(2) ∇f and ∇f^* are inverse of each other. Very interesting fact.
Start from (1). Replace f by f^*, and f^* by $f^{**} = f$.

$$f(x) = \sup_y \left[\langle x, y\rangle - f^*(y)\right], maximized\ when\ x = \nabla f^*(y).$$

$$f(x) = \langle x.y\rangle - f^*(y), x = \nabla f^*(y)$$

From (1), $\langle x, \nabla f(x)\rangle - f^*(\nabla f(x)) = f(x)$. □

2.3 Weak Convergence distances

BL denotes bounded Lipschitz that $||f||_\infty \leq 1, Lip - 1$.

$$\sup_{f \in BL} \left| \int f d\mu - \int f d\nu \right|$$

Consider

$$W_2^2(\mu, \nu) = \inf_{\Pi(\mu,\nu)} \int ||y - x||^2 d\pi = dual\ representation$$

Then we can see Brenier's Theorem.

$$\nabla f : \mu \to \nu$$

$$\nabla f^* : \nu \to \mu$$

3 Kantorovich Duality

3.1 Review of Convex functions

$f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, convex and lower semicontinuous.
We can define dual/conjugate with

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \left[\langle x, y\rangle - f(x)\right]$$

1. $\forall x, y,\ f(x) + f^*(y) - \langle x, y\rangle \geq 0, = 0$ iff $y = \nabla f(x)$ or $x = \nabla f^*(y)$.
2. $\nabla f(\nabla f^*(x)) = x$

Example 3.1. $d = 1$. $f(x) = \begin{cases} x\log x, & x \geq 0 \\ \infty, & x < 0 \end{cases}$. cx lsc.

Check convexity,

$$f'(x) = 1 + \log x$$

Check lsc.

$$\lim_{x\to 0} x\log x = 0$$

Let $y = 1 + \log x$, $x = e^{y-1}$.

$$(f^*)'(y) = e^{y-1}$$

$$f^*(y) = \sup_x [xy - x\log x] = \sup_{x\geq 0} [xy - x\log x] = e^{y-1}$$

Domain of f^* is $\mathbb{R}$ and $Domain(f) = [0, \infty)$.
Another observation
Take f cx and lsc

$$\begin{aligned} \inf_{x\in\mathbb{R}^d} f(x) &= -\sup_{x\in\mathbb{R}^d} [-f(x)] \\ &= -\sup_x [\langle x, 0\rangle - f(x)] \\ &= -f^*(0) \end{aligned}$$

The infimum is attained via checking the dual at 0.
Let x^* is the unique minimizer,

$$\nabla f(x^*) = 0, x^* = \nabla f^*(0)$$

$$x^* = \nabla f^*(0)$$

3.2 Kantorovich Duality

Very similar to 3.1, but in infinity dimension.

Consider the optimal transport problem with a continuous cost $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$.

For $\pi \in \Pi(\mu, \nu)$,

$$I(\pi) = \int c(x, y) d\pi$$

cost of transport using the plan π.

We wish to find out

$$\inf_{\pi \in \Pi} I(\pi)$$

This is doing in the space of functions/measures.

For any function $\varphi \in L^1(\mu)$ and $\psi \in L^1(\nu)$. L^1 means that the integral is finite.

In this case, $\int |\varphi| d\mu < \infty$.

Define

$$J(\varphi, \psi) = \int \varphi(x) d\mu + \int \psi(y) d\nu$$

Let $\Phi = \{\varphi, \psi \textit{ such that } \varphi(x) + \psi(y) \leq c(x, y), \forall x, y\}$.

Theorem 3.2. *(Kantorovich Duality)*

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

The supremum above does not change if we restrict φ, ψ to be bounded continuous functions.

One side is obvious.

Suppose $\pi \in \Pi(\mu, \nu)$. Take any φ, ψ satisfying $\varphi(x) + \psi(y) \leq c(x, y), \forall x, y$.

$$c(x, y) \geq \varphi(x) + \psi(y)$$

$$\begin{aligned} I(\pi) = \int c(x, y) &\geq \int \varphi(x) d\pi + \int \psi(y) d\pi \\ &= \int \varphi(x) d\mu + \int \psi(y) d\nu \\ &\geq \sup_{\Phi} [J(\varphi, \psi)] \end{aligned}$$

Therefore,

$$\inf_{\pi \in \Pi} I(\pi) \geq \sup_{\Phi} [J(\varphi, \psi)]$$

K-duality "=" means there is no duality gap. Minimax inequalities.

3.2.1 Quadratic Cost

$$\begin{aligned} c(x, y) &= \frac{1}{2} ||y - x||^2 \\ &= \frac{1}{x} ||x||^2 + \frac{1}{2} ||y||^2 - \langle x, y \rangle \end{aligned}$$

$$I(\pi) = \int c(x,y)d\pi = \frac{1}{2}\int ||x||^2 d\pi + \frac{1}{2}\int ||y||^2 d\pi - \int \langle x, y\rangle\, d\pi$$
$$= \frac{1}{2}\int ||x||^2 d\mu + \frac{1}{2}\int ||y||^2 d\nu - \int \langle x, y\rangle\, d\pi$$

$$\inf_{\Pi(\mu,\nu)} I(\pi) = \frac{1}{2}\mathbb{E}_\mu ||x||^2 + \frac{1}{2}\mathbb{E}_\nu ||y||^2 - \sup_\pi \int \langle x, y\rangle\, d\pi$$

We give this a name Wasserstein-2 distance between μ and ν that

$$W_2^2(\mu,\nu) = \inf_{\Pi(\mu,\nu)} I(\pi)$$

Fact. *$W_2(\mu,\nu)$ is a metric on $P(\mathbb{R}^d)$ with finite second moment.*

By K-duality,

$$W_2^2(\mu,\nu) = \inf_{\Pi(\mu,\nu)} I(\pi) = \frac{1}{2}\mathbb{E}_\mu ||x||^2 + \frac{1}{2}\mathbb{E}_\nu ||y||^2 - \sup_{\Pi(\mu,\nu)} \int \langle x, y\rangle\, d\pi$$
$$= \sup_\Phi \left[\int \varphi d\mu + \int \psi(y) d\nu\right]$$

$$\sup_{\Pi(\mu,\nu)} \int \langle x, y\rangle\, d\pi = \inf_\Phi \left[\frac{1}{2}\int \left(\underbrace{||x||^2 - \varphi(x)}_{f(x)}\right) d\mu + \frac{1}{2}\int \left(\underbrace{||y||^2 - \psi(y)}_{g(y)}\right) d\nu\right]$$

Constraints here is

$$\varphi(x) + \psi(y) \leq \frac{1}{2}||y-x||^2 = \frac{1}{2}||y||^2 + \frac{1}{2}||x||^2 - \langle x, y\rangle$$

$$f(x) + g(y) \geq \langle x, y\rangle$$

$$\sup_{\Pi(\mu,\nu)} \int \langle x, y\rangle\, d\pi = \inf_{f(x)+g(y)\geq\langle x,y\rangle, \forall x,y} \left[\int f(x)d\mu + \int g(y)d\nu\right]$$

Are their such functions satisfy this constraint??? Yes!
Recall that

$$f(x) + f^*(y) - \langle x, y\rangle \geq 0$$

Now fix f,

$$\inf \int f(y)d\nu, g(y) \geq \langle x, y\rangle - f(x), \forall x$$

$$g(y) \geq \sup_x \left[\langle x, y\rangle - f(x)\right] = f^*(y)$$

Therefore,

$$\inf \int g(y) d\nu = \int f^*(y) d\nu$$

$$\begin{aligned}
\sup_{\Pi(\mu,\nu)} \int \langle x, y\rangle \, d\pi &= \inf_{f(x)+g(y)\geq\langle x,y\rangle, \forall x,y} \left[\int f(x)d\mu + \int g(y)d\nu\right] \\
&= \inf_{f\in L^1(\mu)} \left[\int f(x)d\mu + \int f^*(y)d\nu\right]
\end{aligned}$$

Fix f^*, and we optimize over f.

$$\begin{aligned}
&= \inf_f \left[\int f^{**}(x)d\mu + \int f^*(y)d\nu\right] \\
\sup_\pi \int \langle x, y\rangle \, d\pi &= \inf_{f \ cx,lsc} \left[\int f(x)d\mu + \int f^*(y)d\nu\right]
\end{aligned}$$

This trick called **double convexification trick**.

Ultimate form of W2 distance

$$\frac{1}{2} W_2^2(\mu, \nu) = \frac{1}{2}\left[\int ||x||^2 d\mu + \int ||y||^2 d\nu\right] - \inf_{f \ cx \ lsc.} \left[\int f(x)d\mu + \int f^*(y)d\nu\right]$$

3.2.2 Other cost functions

Earth Move Distance: Wasserstein-1 distance

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

$$W_1(\mu, \nu) = \inf_{\Pi(\mu,\nu)} \int ||y - x|| d\pi$$

What is it dual representation?
K-duality says

$$= \sup_{f \ Lip} \left|\int f d\mu - \int f d\nu\right|$$

Lipschitz-1 means

$$|f(x) - f(y)| \leq ||x - y||$$

Recall There is a metric for weak convergence given by

$$d(\mu,\nu) := \sup_{f\in BL}\left|\int f d\mu - \int f d\nu\right|$$

If we have

$$\lim_{n\to\infty} d(\mu_n,\nu) = 0$$

means that (μ_n) weakly converges to ν.

$$d(\mu,\nu) \leq W_1(\mu,\nu)$$

W_1 gives a stronger topology.
Consider only probability measures that supported on a compact set.
In this case, these topologies are equivalent.

3.2.3 Wasserstein-p Distance

$$W_p(\mu,\nu) = \inf_{\Pi(\mu,\nu)}\int ||x-y||^p d\pi$$

This is Wasserstein p metric. If $p \neq 2$, there is no convenience reformulation of K-duality.

$p = \infty$

$$W_\infty = \inf_{\Pi(\mu,\nu)} \underbrace{ess\sup_{\pi}(||y-x||)}_{\inf\{a>0:\pi(||y-x||\leq a)=1\}}$$

$p = 0$ This is the total variation.

$$c(x,y) = \mathbf{1}\{x\neq y\} = \begin{cases} 1, & if\ x\neq y \\ 0, & otherwise \end{cases}$$

K-duality still holds and can be reformulated as

$$\begin{aligned} ||\mu-\nu||_{TV} &= \inf_{\pi\in\Pi(\mu,\nu)} \pi(x\neq y) \\ &= \sup_{A\ Borel} |\mu(A)-\nu(A)| \end{aligned}$$

Strassen's Theorem (1950).

Proof. Idea of proof of K-duality

$$\inf_{\pi} I(\pi) = \sup_{\Phi}\left[\int \varphi d\mu + \int \psi d\nu\right]$$

Consider indicator function of $\Pi(\mu, \nu)$.

$$M_+ = space\ of\ nonnegative\ measures$$

$$F(\pi) = \begin{cases} 0, & if\ \pi \in \Pi(\mu, \nu) \\ +\infty, & \pi \in M_+, \pi \notin \Pi(\mu, \nu) \end{cases}$$

□

Lemma 3.3. *Here we have*

Proof.

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

□

Proof. Take $\pi \notin \Pi(\mu, \nu)$. Assume $(x, y) \sim \pi$, then $x \sim \mu' \neq \mu$.
There is some φ (bounded cont.) s.t.

$$\int \varphi(x) d\mu > \int \psi(x) d\pi$$

$\lambda > 0$,

$$\lambda[\int \varphi(x) d\mu - \int \varphi(x) d\pi] > 0$$

Let $\lambda \to \infty$. Thus, there exists something let

$$F(\pi) = \infty$$

If $\pi \in \Pi$, we can construct

$$F(\pi) = 0$$

□

Proof. Back to the previous proof

$$I(\pi) + F(\pi) = \inf_{\pi} I(\pi) = \sup_{\Phi} \left[\int \varphi d\mu + \int \psi d\nu \right]$$

$$
\begin{aligned}
\inf_{\Pi} I(\pi) = &\inf_{\pi \in M_+} [I(\pi) + F(\pi)] \\
&= \inf_{M_+} \left[\int c d\pi + \sup_{\varphi,\psi} \left[\int \varphi d\mu + \int \psi d\nu - \int (\varphi + \psi) d\pi \right] \right] \\
&= \inf_{M_+} \sup_{\varphi,\psi} \left[\int \varphi d\mu + \int \psi d\nu - \int (\varphi(x) + \psi(y) - c(x,y)) d\pi \right] \\
&=^{MinMax} \sup_{\varphi,\psi} \left[\int \varphi d\mu + \int \psi d\nu - \sup_{M_+} \int (\varphi(x) + \psi(y) - c(x,y)) d\pi \right] \\
&= \sup_{\varphi,\psi,\varphi(x)+\psi(y) \le c(x,y)} \left[\int \varphi d\mu + \int \psi d\nu \right]
\end{aligned}
$$

□

4 Brenier's Theorem

4.1 Review of duality

$\mu, \nu \in \mathbb{R}^d$, $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$.

$$I(\pi) = \int c(x,y) d\pi$$

$$\inf_{\pi \in \Pi(\rho_0,\rho_1)} I(\pi) = \sup_{\Phi} J(\phi,\psi) = \sup_{\Phi} \int \varphi(x) d\mu(x) + \int \psi(y) d\nu(y)$$

Here, $\varphi(x) + \psi(y) \le c(x,y), \forall x, y$.
$c(x,y) = \frac{1}{2}||y - x||^2$

$$\inf_{\Pi(\mu,\nu)} I(\pi) = \frac{1}{2} W_2^2(\mu,\nu)$$

Duality takes form that

$$\sup_{\pi} \int \langle x, y \rangle \, d\pi(x,y) = \inf_{cx,lsc} \left[\int f(x)\mu(dx) + \int f^*(y)\nu(dy) \right]$$

For

$$\frac{1}{2} W_w^2(\mu,\nu) = \sup_{cx,lsc} \left[\int (\frac{1}{2}||x||^2 - f(x))\mu(dx) + \int \left(\frac{1}{2}||y||^2 - f^*(y) \right) \nu(dy) \right]$$

Transformed functions

$$\phi(x) = \frac{1}{2}||x||^2 - f(x)$$

$$\phi^*(y) = \frac{1}{2}||y||^2 - f^*(x)$$

These functions are c-concave functions and its dual. A pair of dual c-concave.

4.2 Brenier's Theorem

Theorem 4.1. *Let μ, ν be two probability measures with finite second moments. Then, $\exists (f, f^*)$ a pair of cx, lsc function such that*

$$\sup_{\Pi(x,y)} \int \langle x, y \rangle \, d\pi(x,y) = \int f(x)\mu(dx) + \int f^*(x)\nu(dy)$$

Theorem 4.2. *(Breniers' 87) Suppose μ is absolutely continuous. Then,*

1. There is a unique optimal coupling π of the Monge-Kantorovich OT problem given by $(X, \nabla f(X)), X \sim \mu$. Here, ∇f is the unique (uniquely determined μ almost everywhere) gradient of a convex function f such that ∇f pushforwards $\mu$to ν. This function f also attains the maximum in the duality (in Thm. 4.1).

2. ∇f is the unique solution to the Monge problem

$$\int ||x - \nabla f(x)||^2 \mu(dx) = \min_{T_{\#\mu=\nu}} \int ||x - T(x)||^2 \mu(dx)$$

3. Suppose $\nu$is also absolutely continuous. Then, for μa.e. x and νa.e. y,

$$\nabla f \circ \nabla f^*(y) = y, \nabla f^* \circ \nabla f(x) = x$$

Here, ∇f^* is the unique solution to the OT problem transporting μ to ν.

Proof. We already know there is an optimal coupling π,
Duality,

$$\int \langle x, y \rangle \, d\pi^* = \sup_{\Pi(\mu,\nu)} \int \langle x, y \rangle \, d\pi =^{duality} \int f(x)\mu(dx) + \int f^*(y)\nu(dy)$$

$$\int (f(x) + f^*(y) - \langle x, y \rangle) d\pi^*(x,y) = 0$$

Thus, π^* a.e. (x, y), we have

$$f(x) + f^*(y) = \langle x, y \rangle$$

Further, we must have $y = \nabla(x), \forall \mu \; a.e. \; x$.

Thus,

$$\pi^* =^{Law} (x, \nabla f(x))\,, \; for \; f \; that \; attains \; max \; in \; duality.$$

□

This argument is showing that any optimal coupling is given by $\nabla f(x)$, where f attains duality.

Suppose, you found some f such that ∇f pushforward $\mu$to ν.

Can you claim the optimal coupling $\pi^* =^{Law} (x, \nabla f(x))$.

Benefit of duality.

Define $\pi = Law\ of\ (X, \nabla f(x))$.

$$\begin{aligned}\int \langle x, y\rangle \, d\pi &= \int \langle x, \nabla f(x)\rangle \, d\mu \\ &= \int f(x) d\mu + \int f^*(y) d\nu\end{aligned}$$

$$\sup_{\Pi(\mu,\nu)} \int \langle x, y\rangle \, d\pi = \int \langle x, y\rangle \, d\pi = \int f(x)d\mu + \int f^*(y)d\nu = \inf \left[\int g(x)d\mu + \int g^* d\nu \right]$$

Uniqueness in both LHS/RHS.

We have already argues that any optimal π^* must be given by ∇f, for some cx, lsc function f.

Suppose (f, f^*) and (g, g^*) are two pairs of cx, lsc functions that give optimal couplings.

Proof. Call $(f, f^*) = \pi^*$.

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

Because $\pi^* = (X, \nabla f(X))$,

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

Thus,

$$\int (g(x) + g^*(\nabla f(x)) - \langle x, \nabla f(x)\rangle) \mu(dx) = 0$$

Thus,

$$g(x) + g^*(\nabla f(x)) - \langle x, \nabla f(x)\rangle = 0, \; \mu \; a.e.$$

Thus,

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

Uniqueness! □

Solving OT for quadratic cost is equivalent looking for an optimal convex, lsc function.

Theorem 4.3. *Let φ be a cx, lsc function, and let π be a coupling of μ, ν s.t.*

$$\int (\varphi(x) + \varphi^*(y) - \langle x, y \rangle) d\pi(x, y) \leq \epsilon$$

Then,

$$I(\pi) \leq \left(\inf_{\Pi} I \right) + \epsilon = \frac{1}{2} W_2^2(\mu, \nu) + \epsilon.$$

4.3 Cyclical monotonicity

Suppose we have discrete distributions that

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}, \nu = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_j}$$

$$\{x_1, ..., x_N\}, \qquad \{y_1, ..., y_N\}$$

Optimal matching problem. Double stochastic matrices

$$\min_{\Pi(\mu,\nu)} \int ||y - x||^2 d\pi = \min_{\sigma \in S_N} \frac{1}{N} \sum_{i=1}^{N} ||x_i - y_{\sigma_i}||^2$$

$$S_N \textit{ is permutation of } \{1, 2, ..., N\}$$

Question: Can one characterize the set of permutations where the minimum is achieved?

WLOG, assume the identity permutation is optimal.

$$\sum_{i}^{N} ||x_i - y_i||^2 \leq \sum_{i=1}^{N} ||x_i - y_{\sigma_i}||^2, \forall \sigma \in S_N.$$

Consider permutation containing a single non-trivil cycle. One non-trivil cycle, others are identity (single cycle).

[11 10 5 2 1] [3 3] [4 4] [6 6]

[i_1 i_2 i_3...i_m] [3 3] [4 4] [6 6]

Identity elsewhere. To be optimal, we must have

$$\sum_{l=1}^{m} ||x_{i_L} - y_{i_L}||^2 \leq \sum_{l=1}^{m} ||x_{i_L} - y_{i_{L-1}}||^2$$

Definition 4.4. A Set of points $\{(x_1, y_1), ..., (x_N, y_N)\}$ is called cyclically monotone if for all $m \geq 1$, and all cycles $i_1 \leftarrow i_2 \leftarrow i_3 \leftarrow ... \leftarrow i_m$, the following holds.

$$\sum_{l=1}^{m} ||x_{i_L} - y_{i_L}||^2 \leq \sum_{l=1}^{m} ||x_{i_L} - y_{i_{L-1}}||^2$$

Theorem 4.5. *Identity is the optimal permutation if and only if* $\{(x_1, y_1), ..., (x_n, y_n)\}$ *is cyclically monotone.*

Proof. Every permutation can be decomposed as union of disjoint cycles.

If

$$\sum ||x_i - y_i|| \leq \sum ||x_i - y_{\sigma_i}||^2$$

□

Definition 4.6. A subset $\Gamma \subseteq \mathbb{R}^d \times \mathbb{R}^d$ is called cyclically monotone if for any collection of $\{(x_1, y_1), ..., (x_m, y_m)\} \subseteq \Gamma$,

$$\sum ||x_i - y_i||^2 \leq \sum_{i=1}^{m} ||x_i - y_{i-1}||^2$$

Theorem 4.7. *Any optimal coupling π^* of MK OT problem, must be concentrated ($\exists \Gamma \subseteq \mathbb{R}^d \times \mathbb{R}^d$, cyclically monotone, $\pi^*(\Gamma) = 1$) on a cyclically monotone set.*

Proof. Want to couple $\mu$to ν.

We will sample $X_1, ..., X_N \sim \mu$, $y_1, ..., y_N \sim \nu$. Match these optimally.

Have

$$\pi^*_N \to_{N \to \infty} \pi^*$$

from support

$$\Gamma_N \to \Gamma$$

□

4.4 Connect Brenier Theorem to Cyclically monotonicity

If we have μabs. cont., if we have ν

$$\pi^* =^{Law} (X, \nabla f(X)), x \sim \mu$$

$$\Gamma = \left\{(x, \nabla f(x)), x \in \mathbb{R}^d\right\}$$

Rockafeller's Theorem.

If Γ is cyclically monotone, Conversely, any maximumally cyclically monotone subset, must be given by $\left\{(x, \partial f(x)), x \in \mathbb{R}^d\right\}$.

5 Lecture 5

5.1 Review Brenier's Theorem

Example 5.1. $\mathbb{R}^d$, and we have $\mu = \mathcal{N}(a_1, \Sigma_1)$, $\nu = \mathcal{N}(a_2, \Sigma_2)$. What is the optimal MK map between them?

Consider the map $T : \mathbb{R}^d \to \mathbb{R}^d$,

$$T(x) = a_2 + A(x - a_1), A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$$

We can know A is symmetric and PD.

If $X \sim \mathcal{N}(a_1, \Sigma_1)$, then $y = T(x) \sim \mathcal{N}(\cdot, \cdot)$.

$$\mathbb{E}[y] = a_2 + \mathbb{E}[A(x - a_1)] = a_2$$

We know

$$\begin{aligned}\Sigma_y &= A\Sigma_1 A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}\Sigma_1\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2} \\ &= \Sigma_2\end{aligned}$$

How do we know this is a gradient of a convex function?

Define

$$\begin{aligned}f(x) &= \langle a_2, x\rangle + \frac{1}{2}\langle (x - a_1), A(x - a_1)\rangle \\ &= a_2^T x + \frac{1}{2}(x - a_1)^T A(x - a_1)\end{aligned}$$

Then

$$\nabla f(x) = a_2 + A(x - a_1) = T(x)$$

Since $f(x)$, A is PD, we know $f(x)$ is convex.

Then, $T(x)$ is the optimal map.

$$W_2^2(\mu,\nu) = \mathbb{E}_\mu||T(x) - x||^2 = \mathbb{E}||a_2 + A(x - a_1) - x||^2 = ||a_2 - a_1||^2 + tr(\Sigma_z)$$

Consider

$$\begin{aligned} z &= T(x) - x = a_2 + A(x - a_1) - x \sim \mathcal{N}\left[a_2 - a_1, \Sigma_z\right] \\ &= a_2 - Aa_1 + (A - I)x \end{aligned}$$

$$\begin{aligned} \Sigma_z &= (A - I)\Sigma_1(A - I) \\ &= \Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2} \end{aligned}$$

$$W_2^2(\mu,\nu) = ||a_2 - a_1||^2 + Tr\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right]$$

If $a_1 = a_2$,

$$W_2^2(\mu,\nu) = Tr\left[\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right]$$

Bures metric square on PSD matrices.
If $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$, then

$$W_2^2(\mu,\nu) = ||a_1 - a_2||^2 + Tr\left[(\Sigma_1^{1/2} - \Sigma_1^{1/2})^2\right]$$

If $\Sigma_1 = \Sigma_2$, then

$$W_2^2(\mu,\nu) = ||a_1 - a_2||^2$$

Example 5.2. Let $\mu \sim Unif(D), D = \{(x,y) : x^2 + y^2 \leq 1\}$. Let $\nu \sim Unif(U), U = \{(x,y) : x^2 + y^2 = 1\}$.
A natural guess is to do

$$T(x) = \frac{x}{||x||}$$

How do I verify T is optimal?
Consider

$$f(x) = ||x||$$

$$\nabla f(x) = \frac{x}{||x||}$$

Thus, T is optimal for the quadratic cost.

Example 5.3. Take unit square, take μ- Uniform distribution over $[0,1]^2$. And $\nu =$ discrete uniform over $\{(0,0),(1,1),(1,0),(0,1)\}$.

Optimal map for transporting μ to ν?

Another natural guess is to

$$T(x) = (1(x > 1/2), 1(y > 1/2))$$

Convex function

$$f(x) = (x - 1/2)^+ + (y - 1/2)^+$$

$$z^+ = max(z, 0)$$

Twist,

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

5.2 Optimal transport in 1-dimension

$X \sim \mu, y \sim \nu$ on $\mathbb{R}$. Find OT from $\mu$to ν for $c(x,y) = ||y - x||^2$.

Cumulative distribution function (CDF).

$$F_\mu(t) = P(x \le t)$$

F_μ is non-decreasing and

$$\lim_{t \to -\infty} F_\mu(t) = 0$$

$$\lim_{t \to \infty} F_\mu(t) = 1$$

May not be continuous.

Similarly, we could define $F_\nu(t)$.

Lemma 5.4. *Suppose μ is abs. cont. Then define $U = F_\mu(x)$. Then, $U \sim Unif(0,1)$.*

Definition 5.5. Define inverse CDF that

$$F_\mu^{-1}(t) = \inf\{x : F_\mu(x) \ge t\}$$

Corollary 5.6. *We have*

$$\{t \le F_\mu(x)\} = \left\{F_\mu^{-1}(t) \le x\right\}$$

Proof. Pick $0 \le t \le 1$,

$$P(U \ge t) = P_\mu\left(F_\mu(x) \ge t\right) = P_\mu(x \ge F_\mu^{-1}(t)) = 1 - t$$

$\square$

Lemma 5.7. *[Inverse Sampling] Suppose $U \sim Unif(0,1)$. Then, $y = F_\nu^{-1}(U)$. Then, $y \sim \nu$.*

Proof. We have

$$P(Y \leq y) = P(F_\nu^{-1}(U) \leq y) = P(U \leq F_\nu(y)) = F_\nu(y)$$

$\square$

$$X \to_{F_\mu} U \to_{F_\nu^{-1}} y$$

If we have $X \sim \mu$, and

$$F_\nu^{-1} \circ F_\nu(x) \sim \nu$$

If we take $T(x) = F_\nu^{-1} \circ F_\mu(x)$. Then, $T(x) = \nabla f(x), \exists f\ c.x.$
T is an increasing function. Thus, define

$$f(x) = \int_0^x T(y)dy$$

then f is convex.
In 1-d, increasing function $\iff$ derivative of a convex function.
Brenier's Theorem $\iff$ $T(x)$ is the OT map for quadratic cost.

5.2.1 Natural

Suppose F is strictly increasing. F^{-1} is a well-defined strictly increasing map. If we take $0 < p < 1$,

$$F^{-1}(p) = pth\ quantile$$

$$F^{-1}(1/2) = median$$

$$F^{-1}(1/2) = 1st\ quantile$$

$x \mapsto T(x)$ Monotone rearrangements (quantile-quantile maps).
Here,
μ1st quantile, median, pth quantile $\mapsto$ ν1st quantile, median, pth quantile
In 1-d, quadratic cost is not special.

$$c(x,y) = h(x-y), h\ strict\ cx.$$

Then, optimal map is monotone rearrangement.

$$c(x,y) = -h(x-y), h\ strict\ concave.$$

Optimal map is anti-monotone.

$$F_\mu^{-1}(p) \longleftrightarrow F_\nu^{-1}(1-p)$$

Example 5.8. $\mu = \frac{1}{2}Unif(0,1) + \frac{1}{2}Unif(3,4)$. $\nu = Unif(3,5)$.

We could clearly see the monotone transform map as optimal transport map.

$$T(x) = \begin{cases} x+3, & if\ 0 \le x \le 1 \\ x+1, & if\ 3 \le x \le 4 \end{cases}$$

5.3 Knothe-Rosenblatt Transport (KR map)

f, g are densities on $\mathbb{R}^d$. $x = (x_1, ..., x_d) \sim f$, $y = (y_1, ..., y_d) \sim g$.

$$f(x_1, ..., x_d) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|2,1}(x_3|x_2, x_1) \dots f_{d|d-1,...,1}(x_d|x_{d-1}, ..., x_1)$$

$$g(y_1, ..., y_d) = g_1(y_1) g_{2|1}(y_2|y_1) g_{3|2,1}(y_3|y_2, y_1) \dots g_{d|d-1,...,1}(y_d|y_{d-1}, ..., y_1)$$

Let T_1 be the monotone map from $f_1 \to g_1$

$$x_1 \sim f_1$$

$$y_1 = T(x_1) \sim g_1$$

$x_1 = x_1$, $y_1 = T(x_1) = y_1$.

$T_{2|x_1}$ monotone map from $f_{2|1}(\cdot|x_1) \to g_{2|1}(\cdot|y_1 = T(x_1))$.

$$y_2 = T_{2|x_1}(x_2)$$

$$(y_1, y_2) \sim g_1(y) g_{2|1}(y_2|y_1)$$

Inductively, given $x_1, ..., x_{k-1}$ and $y_1, ..., y_{k-1}$.

$T_{k|x_{k-1},...,x_1}$ monotone map $f_{k|k-1,...,1}(\cdot|x_{k-1}, ..., x_1) \to g_{k|k-1,...,1}(\cdot|y_{k-1}, ..., y_1)$.

This defines $(x_1, ..., x_d) \mapsto (y_1, ..., y_d)$. KR-map.

1. Need to know inverses of all conditional.
2. KR map is traingular. To generate y_k, I only need to know $x_1, ..., x_k$.
3. Train neural network to produce an estimate.
4. The order in which $x_1, ..., x_d$ appears matter. $d!$ KR map.
5. No optimality. However,

Consider $c_\epsilon(x, y) = \sum_{i=1}^d \lambda_i(\epsilon)(x_i - y_i)^2$ a weighted quadratic cost.

$\lambda_i(\epsilon) > 0$.

Take f, g, OT w.r.t. e_ϵ. T_ϵ

Theorem 5.9. *Suppose* $k = 1, 2, ..., d-1$

$$\lim_{\epsilon \to 0^+} \frac{\lambda_{k+1}(\epsilon)}{\lambda_k(\epsilon)} = 0.$$

Then, $T_\epsilon \to_{L^2(f)} T(KR\ map)$.

5.4 Dynamical Optimal Transport

$\rho_1, \rho^{'}$ densities on $\mathbb{R}^d$.

We have a Brenier map $x \sim \rho$, $\nabla\psi(x) \sim \rho^{'}$. ψ cx Brenier map.

$$x_t = (1-t)x + t\nabla\psi(x)$$

Call this $T_t(x) = (1-t)x + t\nabla\psi(x)$.

$X \sim \rho$,

$$\rho_t =^{Law} x_t \iff \rho_t = T_{t_\#}\rho$$

Definition 5.10. (McCann's displacement interpolation)

$$(\rho_t, 0 \leq t \leq 1)$$

is called the displacement interpolation between ρ and $\rho^{'}$.

Example 5.11. $\rho \sim \mathcal{N}(0, I)$, $\rho^{'} \sim \mathcal{N}(a, I)$.

$$x \mapsto T(x) = x + a$$

$$X_t = (1-t)x + t(x+a) = x + ta$$

$$X_t \sim \mathcal{N}(ta, I) = \rho_t$$

Example 5.12. $\rho \sim Unif(0,1)^d$, $\rho^{'} \sim Unif[0,r]^d$.

$$\rho_t \sim [0, 1-t+tr]^d$$

$$T(x) = rx = \nabla(\frac{1}{2}r||x||^2)$$

$$\begin{aligned} X_t &= (1-t)X + trX \\ &= (1-t+tr)X \end{aligned}$$

Law of $X_t = Unif[0, 1-t+tr]^d = \rho_t$.