Recap:

Setting: given μ & ν $\rightarrow$ distributions on $\mathbb{R}^d$
we want to transform μ to ν
minimizing some (specified) cost.

Monge: Define a valid transport map: $T_\# \mu = \nu$
(if $X \sim \mu$, then $T(X) \sim \nu$) $\leftarrow$

OT Map:
$$T_0 = \underset{T_\# \mu = \nu}{\arg\min} \int c(x, T(x))\, d\mu(x)$$

Kantorovich: Define valid couplings γ
(joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$ with
first marginal μ & second marginal ν)

OT Coupling:
$$\left[\gamma_0 = \underset{\gamma \in \Gamma_{\mu,\nu}}{\arg\min} \int c(x,y)\, d\gamma(x,y) \right]$$

Wasserstein distances:
$$W_p(\mu,\nu) = \left[\min_{\gamma \in \Gamma_{\mu,\nu}} \int \|x-y\|^p\, d\gamma(x,y) \right]^{1/p}.$$

$\rightarrow$ nice metric on distributions

→ More generally, OT gives a way to lift a metric on some space to a metric on distributions defined on that space.

Brenier's Theorem: Squared Euclidean cost

μ, ν have at least 2 bdd moments

μ is absolutely cont. then:

1. Monge has a solution T_0

$T_0 := \nabla\varphi_0$ $\quad \varphi_0$ is convex.

(ie) T_0 is monotonic.

$\nabla\varphi_0$ is unique μ a.e

2. Kantorovich has a solution supported on the graph of a function

$\gamma_0 := (x, T_0(x))$

3. If ν is also absolutely cont. then

$S_0 := \nabla\varphi_0^*$ is the OT map from ν to μ.

Duality: Discrete case:

$$\min_{\gamma} \sum_{ij} c_{ij}\gamma_{ij} \leftarrow$$

$$\gamma_{ij} \geq 0 \quad \sum_i \gamma_{ij} = 1/n, \quad \sum_j \gamma_{ij} = \frac{1}{n}$$

Dual:

$$\max_{\alpha,\beta} \frac{\sum_{i=1}^{n} \alpha_i}{n} + \frac{\sum_{j=1}^{n} \beta_j}{n} \leftarrow$$

$$\underline{\alpha_i} + \underline{\beta_j} \leq \underline{c_{ij}}$$

(Interpretation as a shipper's problem).

Outline

(1) Duality & Polar Factorization Theorem

(2) Statistical optimal transport
Basic questions

(3) Estimating the OT Map.

Polar Factorization Theorem: (Brenier)

$\rightarrow$ $\|\cdot\|^2$, μ, ν μ a.c., $T_0 = \nabla \varphi_0$.

Any other valid transport map T

$$\boxed{T = T_0 \circ S}$$

S: measure preserving map $\mu \rightarrow \mu$.

$$\boxed{S = Id \quad T = T_0}$$

Matrices: any matrix M

$$\boxed{M = \underset{\text{PSD}}{P}\,\underset{\text{orthonormal}}{O}}$$

$$x \sim N(0, I_d) \qquad \underbrace{\varphi(x) = \frac{x^T P x}{2}}$$

$$Mx \sim N(0, MM^T)$$

$$M = \underset{\text{OT Map}}{P} \cdot O \qquad \underline{T_\sharp(x) = Px}$$

Duality:

$$\left.\max_{\alpha,\beta} \underbrace{\frac{\sum_i \alpha_i}{n}} + \underbrace{\frac{\sum_j \beta_j}{n}}\right]$$

$$\alpha_i + \beta_j \leq c_{ij}$$

→ General Kantorovich dual:

$$\begin{matrix}\int |f|\,d\mu < \infty \\ \int |g|\,d\nu < \infty\end{matrix}\Bigg[\ \max_{\substack{f \in L^1(\mu) \\ g \in L^1(\nu)}} \underbrace{\int f\,d\mu} + \underbrace{\int g\,d\nu}$$

$$\rightarrow f(x) + g(y) \leq c(x,y) \quad \forall (x,y)$$

→ Strong duality: c is lower semi-cont. & bdd below, then strong duality holds.

Structural remarks when $p=1$

1. Fix f, optimize over g.

$$g(y) \leq \underbrace{\inf_x \left[c(x,y) - f(x) \right]} := \underline{\underline{f^c(y)}}$$

Optimal $\underline{\underline{g}}$ is $\underline{\underline{f^c}}$.

2. $p=1$, $\quad \underline{c(x,y) = \|x-y\|}$

$\left\{ \begin{array}{l} \to \text{c-transform of } f \text{ is 1-Lipschitz.} \\ \to \text{c-transform of 1-Lip } f, \ -f. \end{array} \right.$

3. Assume they are true:

$$\to \underbrace{\max_{f \in L^1(\mu)} \int \underset{\text{1-Lipschitz}}{(f^{cc})}\, d\mu + \int \underset{\text{-ve of } f^{cc}}{(f^{cc})^c}\, d\nu.}$$

$$\max_{f \in \text{1-Lipschitz}} \int f\, d\mu - \int f\, d\nu$$

$\hookrightarrow$ Kantorovich–Rubenstein duality.

$$W_1(\mu,\nu) = \sup_{f \in \text{1-Lip}} \left[\int f\, d\mu - \int f\, d\nu \right].$$

W_1 is an Integral Probability Metric.

$$\left[\rho(\mu,\nu) = \sup_{f\in \underbrace{\mathcal{F}}} \left| \int f\,d\mu - \int f\,d\nu \right| \right]$$

TV distance is an IPM

$$TV(\mu,\nu) = \sup_{f \in \underbrace{\{\|f\|_\infty \leq 1\}}} \left| \int f\,d\mu - \int f\,d\nu \right|.$$

Maximum Mean Discrepancy

$$MMD(\mu,\nu) = \sup_{f,\ \underbrace{\|f\|_K \leq 1}} \left| \int f\,d\mu - \int f\,d\nu \right|.$$

RKHS unit ball {not very big}.

Property 1: f.

$$g(y) = \underbrace{\inf_x \left[\|x-y\| - f(x) \right]}.$$

$\inf_x \underbrace{f_x(y)}$ $\swarrow$ 1-Lip.

$\tilde{g}(y) = \|x_0 - y\| - f(x_0) \leftarrow$ 1-Lip

$\tilde{g}(y) - \tilde{g}(y') = \|x_0 - y\| - \|x_0 - y'\|$

$\leq \|y-y'\|.$

Conclusion: $g = f^c$ is 1-Lipschitz.

Property 2: If f is 1-Lip, $\underbrace{(-f)^c = f}$.

$$g(y) = \inf_x \left[\|x-y\| + f(x) \right].$$

$\underbrace{g(y) \leq f(y)}_{\text{take } x=y}$ $\qquad$ $f(y) \leq \|x-y\| + f(x)$

$\underbrace{f(y) \leq g(y).}$

$f = g.$, (ie) $(-f)^c = f.$

Duality for squared Euclidean cost.

Starting pt: $\displaystyle \sup_{\substack{f \in L^1(\mu) \\ g \in L^1(\nu)}} \int f\,d\mu + \int g\,d\nu$

$f(x) + g(y) \leq \|x-y\|^2$

Optimal f & g – Kantorovich potentials.

$$\left.\begin{aligned} f(x) &= \|x\|^2 - 2\varphi(x) \\ g(y) &= \|y\|^2 - 2\psi(y) \end{aligned}\right] \text{ define } \varphi \text{ \& } \psi.$$

$$\sup_{\varphi,\psi} \underbrace{\int \|x\|^2 d\mu} + \underbrace{\int \|y\|^2 d\nu} - 2\int \varphi(x)\,d\mu - 2\int \psi(y)\,d\nu$$

$$\varphi(x) + \psi(y) \geq x^T y.$$

Mod. Dual: $\inf_{\varphi, \psi} \int \varphi \, d\mu + \int \psi \, d\nu$

$$\varphi(x) + \psi(y) \geq x^T y.$$

$$\psi(y) \geq \sup_x \left[x^T y - \varphi(x) \right] := \varphi^*(y)$$

$$\boxed{\inf_{\varphi \in L^1(\mu),\ \varphi \text{ is convex.}} \int \varphi \, d\mu + \int \varphi^* d\nu}$$

semi-dual

If you satisfy condns of Brenier's thm:

$T_0 := \nabla \varphi_0.$

where $\varphi_0 = \arg\inf_{\varphi \in L^1(\mu)} \int \varphi \, d\mu + \int \varphi^* d\nu.$

$$\varphi(x) + \varphi^*(\nabla \varphi(x)) = \langle x, \nabla \varphi(x) \rangle.$$

Statistical Optimal Transport.

→ μ & ν not known.

→ $X_1, \dots, X_n \sim \mu$ & $Y_1, \dots, Y_n \sim \nu$.

μ absolute cont.

Target of inference:

1. Estimate T_0.
   $\hat{T}(x) \approx T_0$?

how many samples do I need?

Applications:

* Generative modeling.
$X_1, \dots X_n \sim N(0, I_d)$
$Y_1, \dots Y_n$ — real images.

* Waddington OT:

cell concentrations | cell concentrations.

$t = 0$ | $t = 1$

"broken correspondence".

OT map is a way to recover correspondence.

2. Estimate $W_2(\mu, \nu)$ or $W_1(\mu, \nu)$.

<u>Hypothesis testing</u>:

$X_1, \dots X_n \sim \mu$

$Y_1, \dots Y_n \sim \nu$

<u>Two-sample</u>: $H_0 : \mu = \nu$

$H_1 : W_p(\mu, \nu) \geq \epsilon$

3. <u>Density estimation</u>: $X_1, \dots X_n \sim \mu$.

construct $\hat{\mu}$, so that $W_p(\hat{\mu}, \mu)$ is small.

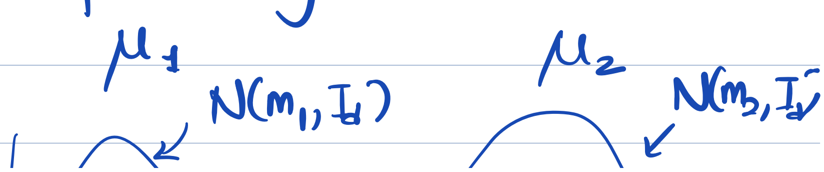
Sample complexity.

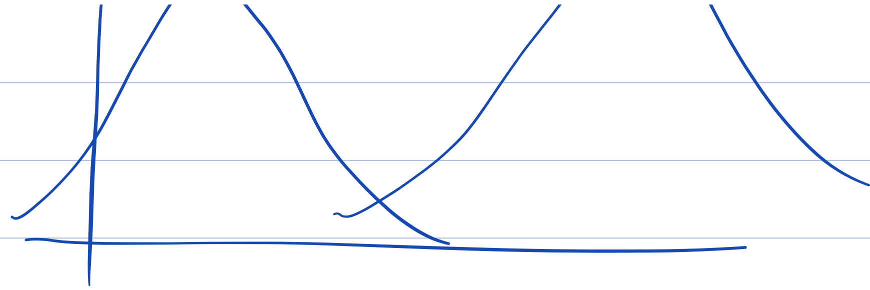
$\rightsquigarrow$ Maybe $\hat{\mu} = \mu_n$ is good / good enough.

4. <u>Barycenter</u>: $\underbrace{\mu_1, \dots \mu_K}$,

$$\underset{\sim}{\bar{\mu}} = \underset{\mu}{\operatorname{argmin}} \sum_i W_2(\mu, \mu_i).$$

"shape preserving".





one defn. of "mid point"

$$= \frac{1}{2} N(m_1, I_d) + \frac{1}{2} N(m_2, I_d)$$

Wasserstein mid-point will be

$$N(\bar{m}, I_d)$$

Estimating the OT map:

Dual Estimators | **Primal Estimators**

Brenier potential

$$\rightarrow \quad \varphi_0 = \underset{\varphi \in \Phi}{\arg\min} \int \varphi \, d\mu + \int \varphi^* \, d\nu$$

$L^1(\mu)$ & convex.

$$T_0 := \nabla \varphi_0 .$$

ERM - empirical risk min.

$$\hat{\varphi} = \underset{\varphi \in \Phi}{\arg\min} \int \varphi \, d\mu_n + \int \varphi^* \, d\nu_n.$$

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \nu_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{y_j}$$

$$\{ \hat{T} = \nabla \hat{\varphi} \}.$$

→ Makkuva et al. (2020)

Input Convex NN (Amos et al.).