

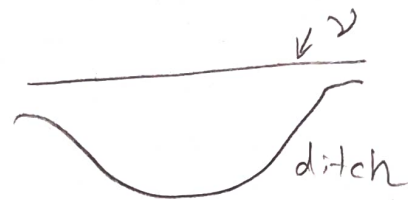
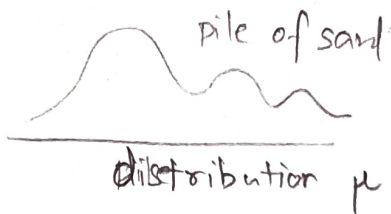
Statistical Optimal Transport

Day-1

Lecture-1

- Topics
- Formulations of OT problem
 - Wasserstein distances
 - Connecting formulation (Brenier's Theorem)
 - Duality.

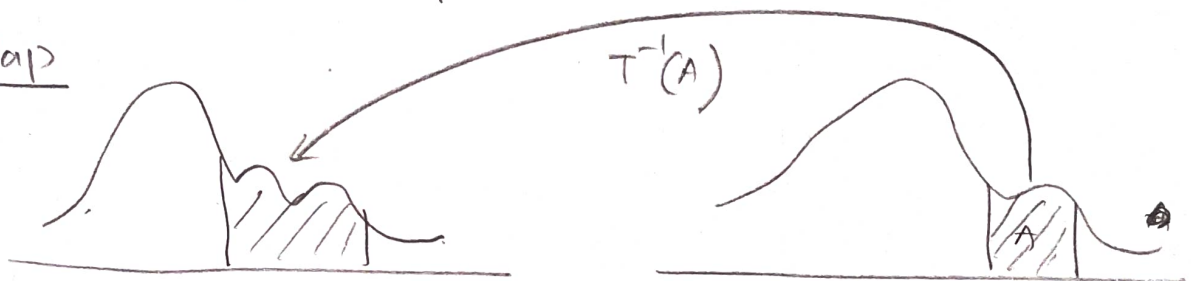
1781 - Monge's Formulation



Transporting : $T : \mathbb{R}^d \rightarrow \mathbb{R}^d ; x \rightarrow T(x)$

⊛ T is the transport map

Valid Map



→ $\mu(T^{-1}(A)) = \nu(A)$ for every set A i.e. $x \sim \mu$ then $T(x) \sim \nu$

→ $C(x, T(x)) = \|x - T(x)\|$
↓ cost.

Monge's Formulation

Find T : $\inf_{\substack{T: \text{valid} \\ T\# \mu = \nu}} \int \|x - T(x)\| d\mu.$

Questions

1. Are there always maps?
2. What can I say about the OT map?
(Stable, regular, unique?)

⊛ Constraint $T_{\#} \mu = \nu$ (non-linear, non-convex constraint)
for all $x: \left\{ \mu(x) = \nu(T(x)) |\det(\nabla T(x))| \right\}$

⊛ There is not always a solution to the problem



⇒ There is no valid map as we need to split atoms

Kantorovich's Formulation (1941)

⊛ Define a coupling of μ & ν . Let γ be the joint distribution of μ & ν . (Thus

$$\forall A \text{ holds: } \gamma(A \times \mathbb{R}^d) = \mu(A) \text{ \& } \gamma(\mathbb{R}^d \times A) = \nu(A)$$

⊛ Denote the set of all couplings as $\mathcal{J}$.

Kantorovich's program:

$$\text{Find } \mathcal{J} \text{ which } \inf_{\gamma \in \mathcal{J}_{\mu, \nu}} \int \|x - y\| d\gamma(x, y)$$

Remarks 1. Constraints set is nice — infinite dimensional linear program.

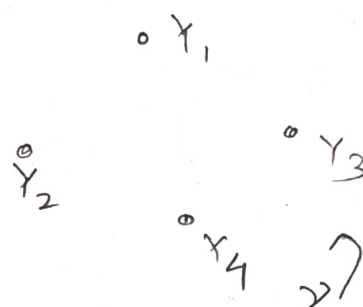
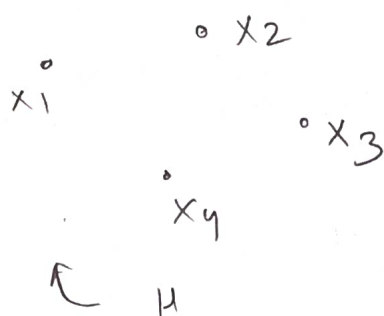
$$\int_{\mathbb{R}^d} \gamma(x, y) dy = \mu(x) \text{ \& } \int_{\mathbb{R}^d} \gamma(x, y) dx = \nu(y)$$

2. There are always valid couplings. $\rightarrow x \sim \mu$ & $y \sim \nu$ sampled independently. Thus independence coupling exists.

3. Kantorovich's coupling gives a Markov Kernel for each x in X . It gives an conditional distribution over where to move mass.

Discrete Case

$$\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \gamma = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$$



δ = matrix size $(n \times n)$

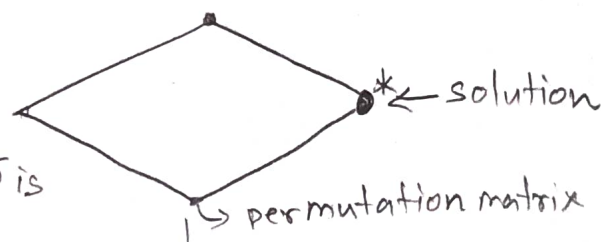
Constraints (a) $\delta_{ij} \geq 0$

marginal constraints (b) $\sum_j \delta_{ij} = \frac{1}{n}$ & $\sum_i \delta_{ij} = \frac{1}{n} \quad \forall i, j$

objective: $C(x, y) = \|x - y\|$

$$\min \sum_{i,j} \delta_{ij} c(x_i, y_j)$$

s.t. $\delta \in \mathcal{J}$



- (*) It defines a polytope i.e. $\mathcal{J}$ is a polytope.
- (*) Birkhoff polytope: convex hull of the set of permutation matrices $\rightarrow$ It has a solution at one end point

(*) Monge's problem also has solution in this case.

Wasserstein Distance

(*) Kantorovich's program with L_p cost

$$C(x, y) = \|x - y\|_p^p$$

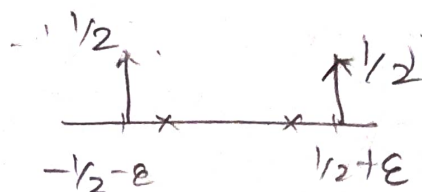
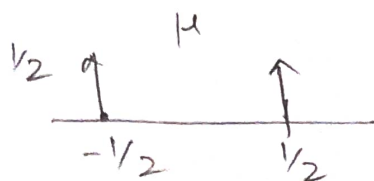
(*) For μ, γ with $\int \|x\|^p d\mu < \infty$ & $\int \|y\|^p d\gamma < \infty$ then

Wasserstein distance $\rightarrow W_p^p(\mu, \gamma) = \left[\inf_{\delta \in \mathcal{J}_{\mu, \gamma}} \int \|x - y\|_p^p d\delta \right]^{1/p}$

(*) Wasserstein distance is cost of the best coupling.

Remarks

1. Metric on distributions
2. Respects underlying metric structure



- (*) Total variation / χ^2 / Hellinger
 $\downarrow$
1 $\downarrow$ ∞ $\rightarrow$ These doesn't care about the structure just measure how much mass at each location
- (*) $W_2^2(\mu, \nu) \leq \varepsilon^2$ in this case $\rightarrow$ Care about structure

Monge's Problem & Kantorovich's Problem Connection

- (*) $p=2$ cost is $c(x, y) = \|x - y\|_2^2$
- (*) Brenier's Theorem: Consider μ, ν with finite 2nd moments & μ has a density w.r.t Lebesgue measure then
1. $\exists$ a map T_0 with $T_0 \# \mu = \nu \rightarrow T_0$ is optimal for Monge's program
- & $T_0 = \nabla \psi_0$ where ψ_0 is convex & unique solution μ almost everywhere. ψ_0 is cyclical monotone.
 2. ϕ_0 which solves Kantorovich's Problem
 $\phi_0 = \inf \int c(x, y) d\gamma$
 $\phi_0 = \text{dist}(x, \nabla \psi_0(x))$
 3. If ν also has a density then
 $S_0 \# \nu = \mu$ & $S_0 = \nabla \psi_0^*$
- (*) Here $\psi_0^*(y) = \sup_x [\langle x, y \rangle - \psi_0(x)]$

Why is $T_0 = \nabla \psi_0$?

(*) Gradient of a convex $\approx$ "multivariate monotone" $\rightarrow$ Why?

Defn 1D Convex function

$$\psi(y) \geq \psi(x) + \nabla \psi(x)^T (y - x)$$

Why monotonicity of convex gradient?

also $\psi(x) \geq \psi(y) + \nabla \psi(y)^T (x-y)$

Thus, $(\nabla \psi(x) - \nabla \psi(y))(x-y) \geq 0$

If $x \geq y$ then $\nabla \psi(x) \geq \nabla \psi(y)$

$\Rightarrow \nabla \phi$ is monotonic

Cyclical monotonicity (Multivariate ^{analog} ~~formation~~ of monotonicity).

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \dots \rightarrow x_k \rightarrow x_1$$

$$\sum_{i=1}^k (x_i - x_{i+1})^T \nabla \phi(x_i) \geq 0$$

A cyclical monotone function lives in the gradient of convex function

Duality (Kantorovich)

Discrete $\min \sum_{i,j} x_{ij} c_{ij}$

$$\begin{aligned} \text{s.t. } & \sum_j x_{ij} = \frac{1}{n} \quad \forall j \rightarrow \alpha_j \\ & \sum_i x_{ij} = \frac{1}{n} \quad \forall i \rightarrow \beta_i \\ & x_{ij} \geq 0 \quad \forall i,j \rightarrow \delta_{ij} \end{aligned} \rightarrow \text{Lagrange Multiplier}$$

Lagrangian Dual

$$\begin{aligned} \min \max_{\alpha, \beta, \delta \geq 0} & \sum_{i,j} x_{ij} c_{ij} - \sum_j \alpha_j \left(\sum_i x_{ij} - \frac{1}{n} \right) \\ & - \sum_i \beta_i \left(\sum_j x_{ij} - \frac{1}{n} \right) - \sum_{i,j} \delta_{ij} x_{ij} \end{aligned}$$

constraints $x_{ij} - \alpha_j - \beta_i - \delta_{ij} = 0$

objective: $\max_{\alpha, \beta} \frac{\sum_j \alpha_j}{n} + \frac{\sum_i \beta_i}{n}$

$\alpha_j + \beta_i \leq c_{ij}$

(*) This is ^{similar} called "Shipper's problem"

α_j to unload in y_j

β_i to load in x_i

Shipper will solve dual.
 (1) clever shipper can get full amount
 ~ strong duality