Optimal Transport: A Crash Course

Gustavo K. Rohde*, Shiying Li *, Soheil Kolouri†

*University of Virginia, †Vanderbilt University

June 1, 2022

Overview

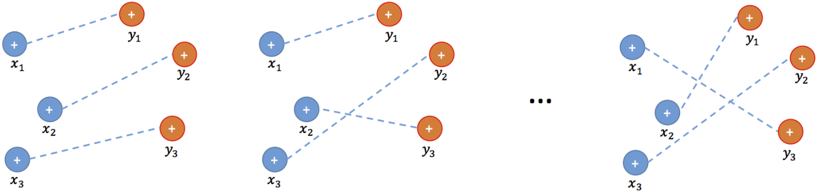
References:

- Kolouri, Park, Thorpe, Slepcev, Rohde, Transport-based analysis, modeling and learning from signal and data distributions. ArXiv, 2017.
- Kolouri, Park, Thorpe, Slepcev, Rohde, Optimal Mass Transport: Signal processing and machine learning applications. IEEE Signal Processing Magazine, 2017.
- Mathew Thorpe, Introduction to Optimal Transport. Preprint, 2018.

Introduction

What is Optimal Transport?

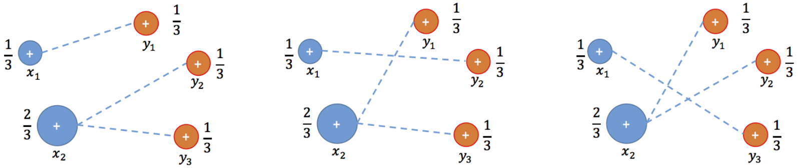
- The optimal transport problem seeks the most efficient way of transporting one distribution of mass into another.



$$\begin{array}{c|ccc} & y_1 & y_2 & y_3 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 0 & 1 & 0 \\ x_3 & 0 & 0 & 1 \end{array} \qquad \begin{array}{c|ccc} & y_1 & y_2 & y_3 \\ \hline x_1 & 1 & 0 & 0 \\ x_2 & 0 & 0 & 1 \\ x_3 & 0 & 1 & 0 \end{array} \qquad \cdots \qquad \begin{array}{c|ccc} & y_1 & y_2 & y_3 \\ \hline x_1 & 0 & 1 & 0 \\ x_2 & 0 & 0 & 1 \\ x_3 & 1 & 0 & 0 \end{array}$$

What is Optimal Transport?

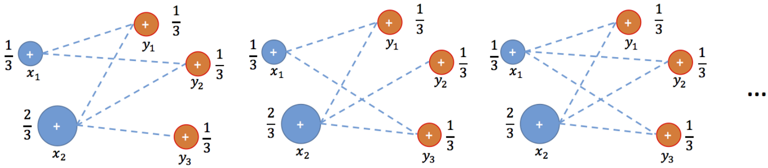
- The optimal transport problem seeks the most efficient way of transporting one distribution of mass into another.



$$\begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{array}
\qquad
\begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} \end{bmatrix} \end{array}
\qquad
\begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix} \end{array}$$

What is Optimal Transport?

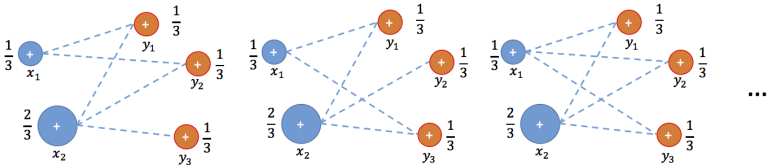
- The optimal transport problem seeks the most efficient way of transporting one distribution of mass into another.



$$\begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{2}{9} & \frac{1}{9} & 0 \\ \frac{1}{9} & \frac{2}{9} & \frac{1}{3} \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \end{bmatrix} \end{array} \qquad \ldots$$

What is Optimal Transport?

- The optimal transport problem seeks the most efficient way of transporting one distribution of mass into another.



$$\begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{ccc} y_1 & y_2 & y_3 \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} \frac{2}{9} & \frac{1}{9} & 0 \\ \frac{1}{9} & \frac{2}{9} & \frac{1}{3} \end{bmatrix} \end{array} \quad \begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \end{array} \quad \begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \end{bmatrix} \end{array} \quad \cdots$$

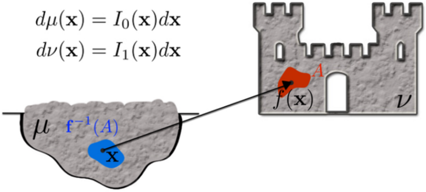
There are infinitely many transportation plans!

A little bit of history!

- The problem was originally studied by Gaspard Monge in the 18'th century.

Gaspard Monge
1746-1818

$$d\mu(\mathbf{x}) = I_0(\mathbf{x})d\mathbf{x}$$
$$d\nu(\mathbf{x}) = I_1(\mathbf{x})d\mathbf{x}$$

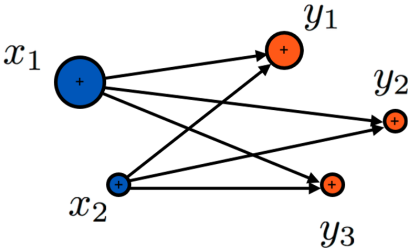


Le mémoire sur les déblais et les remblais
(The note on land excavation and infill)

A little bit of history!:

- Working on optimal allocation of scarce resources during World War II, Kantorovich revisited the optimal transport problem in 1942.

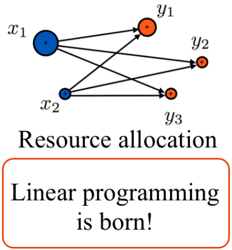
Leonid Kantorovich
1912-1986



Resource allocation

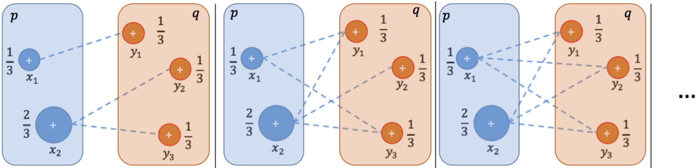
A little bit of history!

- In 1975, Kantorovich shared the Nobel Memorial Prize in Economic Sciences with Tjalling Koopmans "for their contributions to the theory of optimum allocation of resources."



Kantorovich Formulation

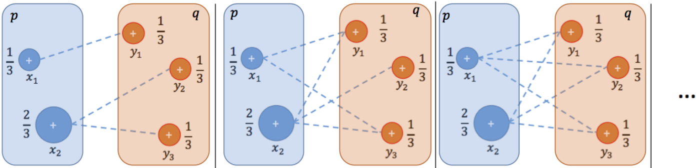
- First lets focus on the common trait of these transportation plans.



$$\gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \ldots$$

$$\sum_i \gamma_{i\cdot} = [\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}] \qquad \sum_j \gamma_{\cdot j} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

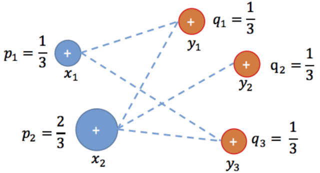
- First lets focus on the common trait of these transportation plans.



$$\gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array} \qquad \ldots$$

$$\sum_i \gamma_{i\cdot} = [\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}] \qquad \sum_j \gamma_{\cdot j} = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}$$

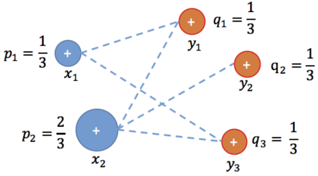
A transportation plan is a joint probability distribution with its marginals equal to the original distributions, p and q.



$$\gamma = \begin{array}{cc} & \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ & \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \begin{array}{c} x_1 \\ x_2 \end{array} \end{array}$$

$$\sum_i \gamma_{ij} = q_j, \sum_j \gamma_{ij} = p_i$$

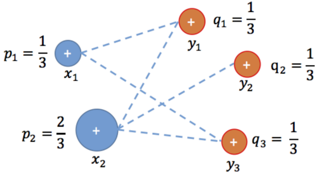
- Let $\mu = \sum_i p_i \delta_{x_i}$ and $\nu = \sum_j q_j \delta_{y_j}$ represent the mass distributions, where δ_{x_i} is a Dirac measure centered at x_i, and $\sum_i p_i = \sum_j q_j = 1$.



$$\gamma = \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \end{array} \begin{array}{c} x_1 \\ x_2 \end{array}$$

$$\sum_i \gamma_{ij} = q_j, \sum_j \gamma_{ij} = p_i$$

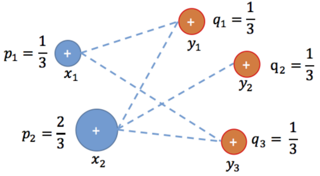
- Let $\mu = \sum_i p_i \delta_{x_i}$ and $\nu = \sum_j q_j \delta_{y_j}$ represent the mass distributions, where δ_{x_i} is a Dirac measure centered at x_i, and $\sum_i p_i = \sum_j q_j = 1$.
- As we mentioned γ_{ij} identifies the amount of mass that is being transported from x_i to y_j.



$$\gamma = \begin{array}{c c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} & \\ \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} & \begin{array}{c} x_1 \\ x_2 \end{array} \end{array}$$

$$\sum_i \gamma_{ij} = q_j, \sum_j \gamma_{ij} = p_i$$

- Let $\mu = \sum_i p_i \delta_{x_i}$ and $\nu = \sum_j q_j \delta_{y_j}$ represent the mass distributions, where δ_{x_i} is a Dirac measure centered at x_i, and $\sum_i p_i = \sum_j q_j = 1$.
- As we mentioned γ_{ij} identifies the amount of mass that is being transported from x_i to y_j.
- Transportation from x_i to y_j would induce a cost $c_{ij} = c(x_i, y_j)$ (e.g. cost of gas for transportation distance)



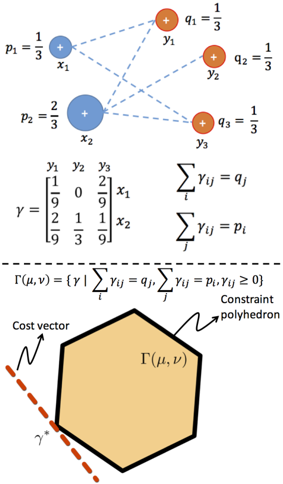
$$\gamma = \begin{array}{c} \\ \\ \end{array}\begin{array}{ccc} y_1 & y_2 & y_3 \end{array}$$

$$\gamma = \begin{bmatrix} \frac{1}{9} & 0 & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \begin{array}{c} x_1 \\ x_2 \end{array}$$

$$\sum_i \gamma_{ij} = q_j, \sum_j \gamma_{ij} = p_i$$

- Let $\mu = \sum_i p_i \delta_{x_i}$ and $\nu = \sum_j q_j \delta_{y_j}$ represent the mass distributions, where δ_{x_i} is a Dirac measure centered at x_i, and $\sum_i p_i = \sum_j q_j = 1$.
- As we mentioned γ_{ij} identifies the amount of mass that is being transported from x_i to y_j.
- Transportation from x_i to y_j would induce a cost $c_{ij} = c(x_i, y_j)$ (e.g. cost of gas for transportation distance)
- Optimal transport problem seeks the most efficient transportation plan with respect to the cost c:

$$\min_\gamma \sum_i \sum_j c_{ij}\gamma_{ij}$$
$$s.t. \quad \sum_i \gamma_{ij} = q_j, \quad \sum_j \gamma_{ij} = p_i,$$
$$\gamma_{ij} \geq 0$$

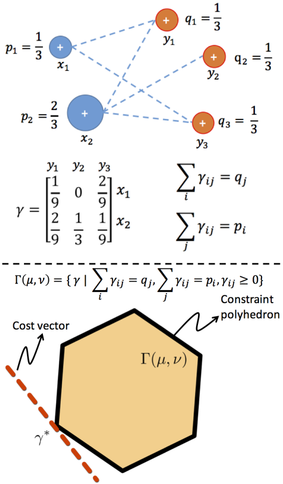


Optimal transport problem:

$$\min_\gamma \sum_i \sum_j c_{ij}\gamma_{ij}$$

$$s.t. \quad \sum_i \gamma_{ij} = q_j, \quad \sum_j \gamma_{ij} = p_i,$$

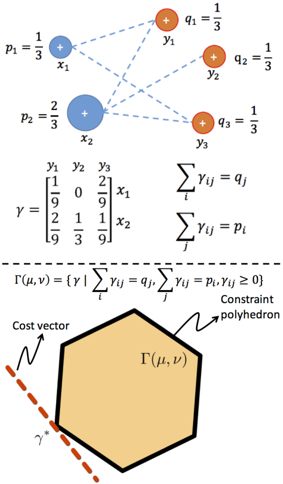
$$\gamma_{ij} \geq 0$$



Optimal transport problem:

$$\min_\gamma \sum_i \sum_j c_{ij}\gamma_{ij}$$
$$s.t. \quad \sum_i \gamma_{ij} = q_j, \quad \sum_j \gamma_{ij} = p_i,$$
$$\gamma_{ij} \geq 0$$

- OT formulation for discrete mass distributions (point cloud distributions) is a linear programing problem



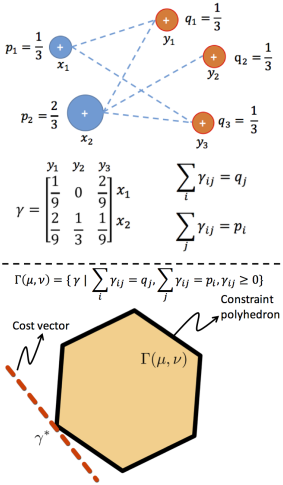
Optimal transport problem:

$$\min_\gamma \sum_i \sum_j c_{ij}\gamma_{ij}$$

$$s.t. \quad \sum_i \gamma_{ij} = q_j, \quad \sum_j \gamma_{ij} = p_i,$$

$$\gamma_{ij} \geq 0$$

- OT formulation for discrete mass distributions (point cloud distributions) is a linear programing problem
- The problem **is convex** but **not strictly convex**.



Optimal transport problem:

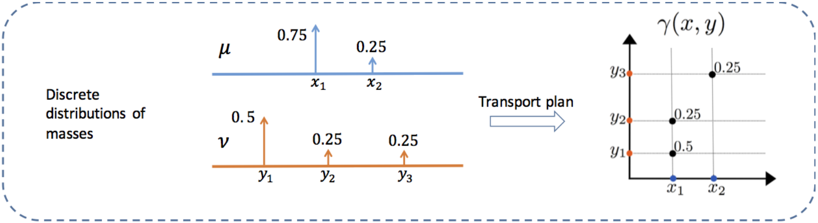
$$\min_\gamma \sum_i \sum_j c_{ij}\gamma_{ij}$$

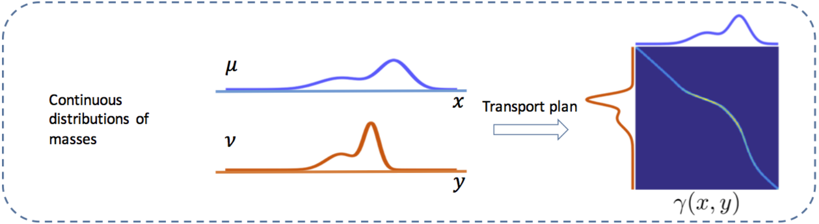
$$s.t. \quad \sum_i \gamma_{ij} = q_j, \quad \sum_j \gamma_{ij} = p_i,$$

$$\gamma_{ij} \geq 0$$

- OT formulation for discrete mass distributions (point cloud distributions) is a linear programing problem
- The problem **is convex** but **not strictly convex**.
- Common solvers include: Simplex algorithm, Interior point methods (AKA Barrier methods), etc.

- What if we have two continuums of masses?





Kantorovich general formulation:

- A transport plan between measures μ and ν defined on X and Y is a probability measure $\gamma \in X \times Y$ with marginals,

$$\gamma(X, A) = \nu(A), \; \gamma(B, Y) = \mu(B)$$

Kantorovich general formulation:

- A transport plan between measures μ and ν defined on X and Y is a probability measure $\gamma \in X \times Y$ with marginals,

$$\gamma(X, A) = \nu(A),\ \gamma(B, Y) = \mu(B)$$

- Let $c(\cdot, \cdot) : X \times Y \to \mathbb{R}$ define the transportation cost from X to Y.

Kantorovich general formulation:

- A transport plan between measures μ and ν defined on X and Y is a probability measure $\gamma \in X \times Y$ with marginals,

$$\gamma(X, A) = \nu(A),\ \ \gamma(B, Y) = \mu(B)$$

- Let $c(\cdot, \cdot) : X \times Y \rightarrow \mathbb{R}$ define the transportation cost from X to Y.
- The transport problem is then formulated as finding the transport plan that minimizes the expected cost, c, with respect to the joint probability measure γ,

$$\begin{aligned} KP(\mu, \nu) &= min_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x, y) d\gamma(x, y) \\ \Gamma(\mu, \nu) &= \{\gamma \mid \gamma(A, Y) = \mu(A),\ \ \gamma(X, B) = \nu(B)\} \end{aligned}$$

Kantorovich: discrete formulation (Earth Mover's Distance)

- Let $\mu = \sum_{i=1}^{N} p_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{M} q_j \delta_{y_j}$, where δ_{x_i} is a Dirac measure,

$$\begin{aligned} KP(\mu,\nu) &= \min_\gamma \sum_i \sum_j c(x_i, y_j)\gamma_{ij} \\ s.t. &\quad \sum_j \gamma_{ij} = p_i, \ \sum_i \gamma_{ij} = q_j, \ \gamma_{ij} \geq 0 \end{aligned}$$

Kantorovich: general formulation

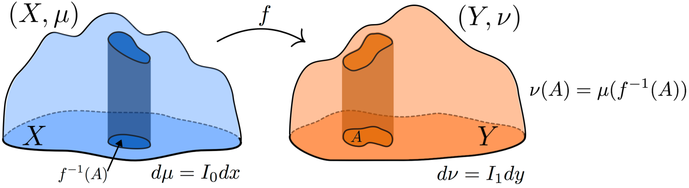
- Let $d\mu(x) = p(x)dx$ and $d\nu(x) = q(x)dx$,

$$\begin{aligned} KP(\mu,\nu) = \quad & \min_\gamma \int_{X\times Y} c(x,y)d\gamma(x,y) \\ s.t. \quad & \int_Y d\gamma(x,y) = p(x), \ \int_X d\gamma(x,y) = q(y) \\ & \gamma(x,y) \geq 0 \end{aligned}$$

Monge Formulation

Monge formulation and Transport Maps:

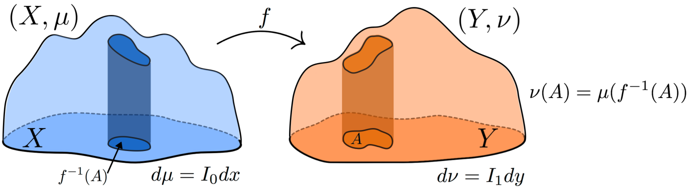
- A map, $f : X \to Y$, for measures μ and ν defined on spaces X and Y and with corresponding densities I_0 and I_1, is called a transport map (or a mass preserving map) iff,



$$MP := \{f \mid \int_{f^{-1}(A)} I_0(x)dx = \int_A I_1(y)dy\}$$

Monge formulation and Transport Maps:

- A map, $f: X \to Y$, for measures μ and ν defined on spaces X and Y and with corresponding densities I_0 and I_1, is called a transport map (or a mass preserving map) iff,



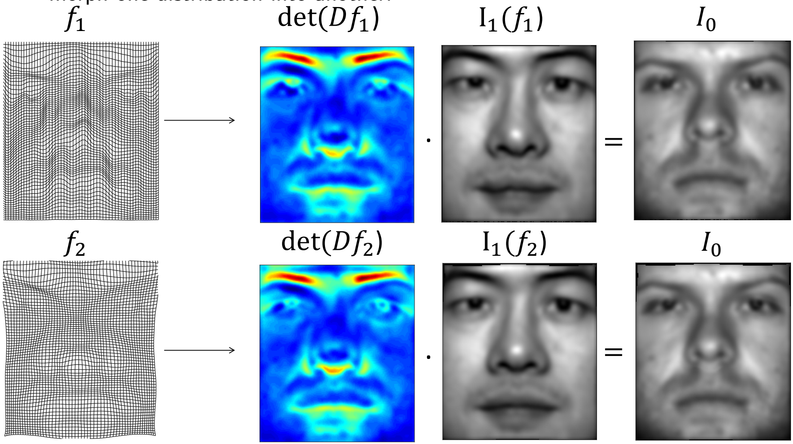
$$MP := \{f \mid \int_{f^{-1}(A)} I_0(x)dx = \int_A I_1(y)dy\}$$

- When f exists and it is differentiable, above constraint can be written in differential form as,

$$MP = \{f : X \to Y | det(Df(x)) I_1(f(x)) = I_0(x), \forall x \in X\}$$

Non-uniqueness of transport maps:

- Similar to transport plans, there exists infinitely many transport maps that morph one distribution into another.



Monge formulation and Transport Maps:

- Find the optimal transport map $f : X \to Y$ that minimizes the expected cost of transportation,

$$M(\mu, \nu) = \inf_{f \in MP} \int_X c(x, f(x)) I_0(x) dx$$

Monge formulation and Transport Maps:

- Find the optimal transport map $f : X \to Y$ that minimizes the expected cost of transportation,

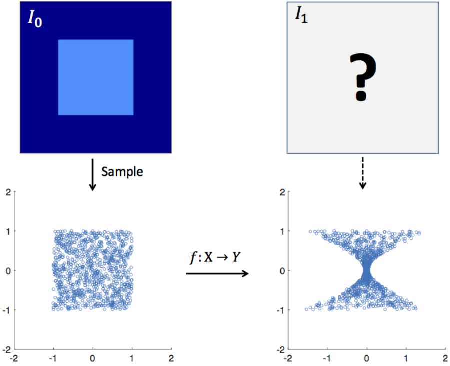
$$M(\mu, \nu) = \inf_{f \in MP} \int_X c(x, f(x)) I_0(x) dx$$

- In the majority of engineering applications the cost is the Euclidean distance,

$$\begin{aligned} M(\mu, \nu) &= \inf_{f \in MP} \int_X |x - f(x)|^2 I_0(x) dx \\ MP &= \{f : X \to Y | \int_{f^{-1}(A)} I_0(x) dx = \int_A I_1(y) dy\} \end{aligned} \tag{1}$$

Note that as opposed to the Kantorovich formulation the objective function and the constraint in Eq. (1) are both nonlinear with respect to f.

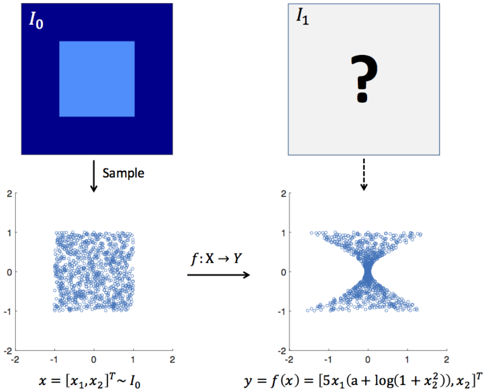
Elucidating Example:



$x = [x_1, x_2]^T \sim I_0$

$y = f(x) = [5x_1(a + \log(1 + x_2^2)), x_2]^T$

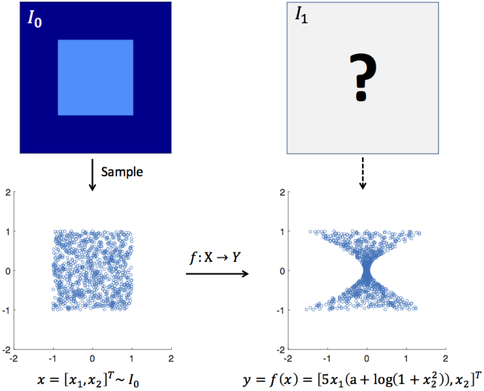
Elucidating Example:



- The distribution of transformed samples follows:

$$I_1(y) = \int_X I_0(x)\delta(y - f(x))dx$$

Elucidating Example:



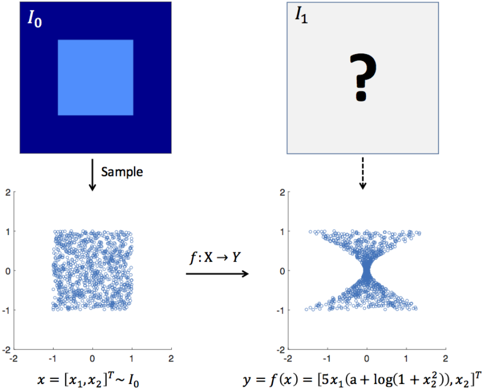
- The distribution of transformed samples follows:

$$I_1(y) \quad = \quad \int_X I_0(x)\delta(y - f(x))dx$$

- When f is a diffeomorphism above equation simplifies to:

$$I_1(y) = det(Df^{-1}(y))I_0(f^{-1}(y))$$

Elucidating Example:

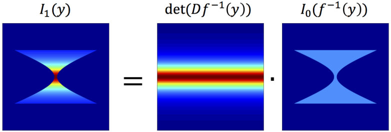


- The distribution of transformed samples follows:

$$I_1(y) \quad = \quad \int_X I_0(x)\delta(y - f(x))dx$$

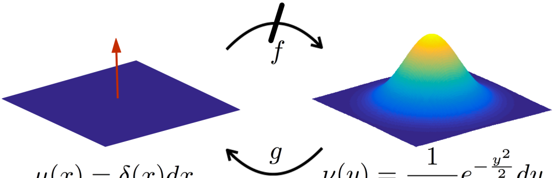
- When f is a diffeomorphism above equation simplifies to:

$$I_1(y) = det(Df^{-1}(y))I_0(f^{-1}(y))$$



A Transport Map May Not Exist:

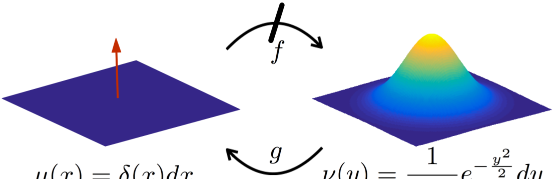
- A transport map, f, exists only if μ is an absolutely continuous measure and $c(x, f(x))$ is convex.
- Here is an example where the transport map does not exists:



$$\mu(x) = \delta(x)dx \qquad \nu(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

A Transport Map May Not Exist:

- A transport map, f, exists only if μ is an absolutely continuous measure and $c(x, f(x))$ is convex.
- Here is an example where the transport map does not exists:



$$\mu(x) = \delta(x)dx \qquad \nu(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

- Monge formulation is not suitable for analyzing point cloud distributions or any particle like distributions.

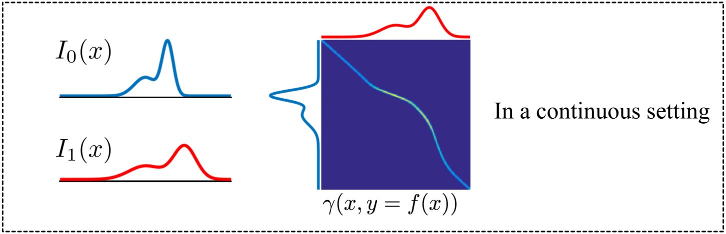
Kantorovich vs. Monge

- The following relationship holds between Monge's and Kantorovich's formulation,

$$KP(\mu,\nu) \leq M(\mu,\nu)$$

- When an optimal transport map exists, $f : X \to Y$, the optimal transport plan and the optimal transport map are related through,

$$\int_{X\times Y} c(x,y)d\gamma(x,y) = \int_X c(x,f(x))d\mu(x)$$



Existence and uniqueness

Brenier's theorem

- Let $c(x,y) = |x-y|^2$ and let μ be absolutely continuous with respect to the Lebesgue measure. Then, there exists a unique optimal transport map $f : X \to Y$ such that,

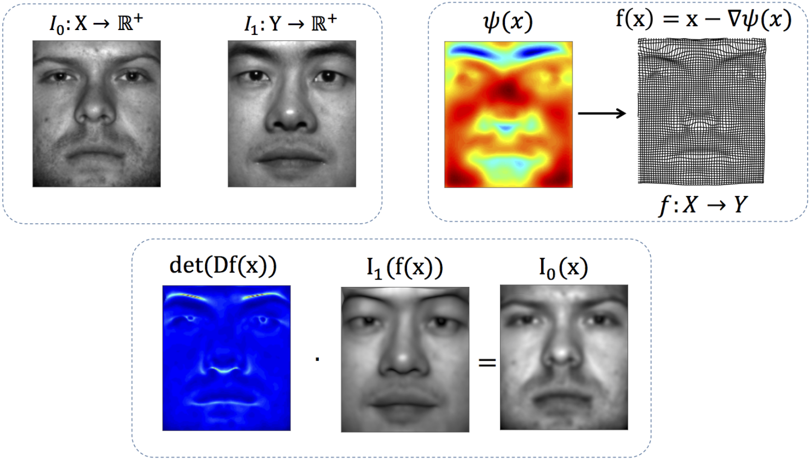
$$\int_{f^{-1}(A)} d\mu(x) = \int_A d\nu(y)$$

which is characterized as,

$$f(x) = x - \nabla\psi(x) = \nabla \underbrace{\left(\frac{1}{2}|x|^2 - \psi(x)\right)}_{\phi(x)}$$

for some concave scalar function ψ. In other words, f is the gradient of a convex scalar function ϕ, and therefore it is curl free.

Implications of the Brenier's theorem



Dual Problem

Kantorovich problem and its dual

- Primal problem:

$$
\begin{aligned}
KP(\mu,\nu) = \quad & \min_\gamma \quad \int_{X\times Y} c(x,y)d\gamma(x,y) \\
& s.t. \quad \int_Y d\gamma(x,y) = p(x), \quad \int_X d\gamma(x,y) = q(y) \\
& \qquad \gamma(x,y) \geq 0
\end{aligned}
$$

Kantorovich problem and its dual

- Primal problem:

$$\begin{aligned} KP(\mu,\nu) = \quad & \min_{\gamma} \quad \int_{X\times Y} c(x,y)d\gamma(x,y) \\ & s.t. \quad \int_{Y} d\gamma(x,y) = p(x), \;\; \int_{X} d\gamma(x,y) = q(y) \\ & \qquad \gamma(x,y) \geq 0 \end{aligned}$$

- Dual problem:

$$\begin{aligned} DP(\mu,\nu) = \quad & \max_{\phi,\psi} \quad \int_{X} \phi(x)d\mu(x) + \int_{Y} \psi(y)d\nu(y) \\ & s.t. \qquad \phi(x) + \psi(y) \leq c(x,y), \quad \forall (x,y) \in X\times Y \end{aligned}$$

Dual problem and Kantorovich-Rubinstein theorem:

- Dual problem:

$$\begin{aligned} DP(\mu,\nu) = \quad & \max_\phi \quad \int_X \phi(x)d\mu(x) + \int_Y \phi^c(y)d\nu(y) \\ & s.t. \qquad \phi^c(y) = \inf_X c(x,y) - \phi(x) \end{aligned}$$

Dual problem and Kantorovich-Rubinstein theorem:

- Dual problem:

$$\begin{aligned} DP(\mu,\nu) = \quad & \max_{\phi} \quad \int_X \phi(x)d\mu(x) + \int_Y \phi^c(y)d\nu(y) \\ & s.t. \qquad \phi^c(y) = \inf_X c(x,y) - \phi(x) \end{aligned}$$

Kantorovich-Rubinstein theorem

- Let μ and ν be two probability measures in the metric space (X, d).
- When the cost function is the ℓ_1 norm, $c(x, y) = |x - y|$, the Dual problem could be simplified into:

$$DP(\mu,\nu) = \quad \max_{\phi\in Lip_1(X)} \quad \int_X \phi(x)d\mu(x) - \int_X \phi(y)d\nu(y)$$

where $Lip_1(X) = \{\phi \ \mid \ |\phi(x) - \phi(y)| \leq d(x,y), \forall x, y \in X\}$.

Transport-Based Metrics

p-Wasserstein distance

- Let $P_p(\Omega)$ be the set of Borel probability measures with finite p'th moment defined on a given metric space (Ω, d). The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$ is then defined as the optimal transport problem with the cost function $c(x, y) = d^p(x, y)$. Let μ and ν be in $P_p(\Omega)$, then,

$$W_p(\mu, \nu) = \left(\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} d^p(x, y) d\gamma(x, y) \right)^{\frac{1}{p}}$$

p-Wasserstein distance

- Let $P_p(\Omega)$ be the set of Borel probability measures with finite p'th moment defined on a given metric space (Ω, d). The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$ is then defined as the optimal transport problem with the cost function $c(x,y) = d^p(x,y)$. Let μ and ν be in $P_p(\Omega)$, then,

$$W_p(\mu,\nu) = \left(\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} d^p(x,y) d\gamma(x,y) \right)^{\frac{1}{p}}$$

or equivalently when the optimal transport map, f^*, exists,

$$W_p(\mu,\nu) = \left(\min_{f \in MP} \int_{\Omega} d^p(x, f(x)) d\mu(x) \right)^{\frac{1}{p}} .$$

p-Wasserstein distance

- Let $P_p(\Omega)$ be the set of Borel probability measures with finite p'th moment defined on a given metric space (Ω, d). The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$ is then defined as the optimal transport problem with the cost function $c(x,y) = d^p(x,y)$. Let μ and ν be in $P_p(\Omega)$, then,

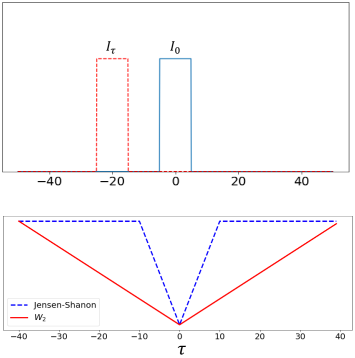
$$W_p(\mu,\nu) = \left(\min_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega \times \Omega} d^p(x,y) d\gamma(x,y) \right)^{\frac{1}{p}}$$

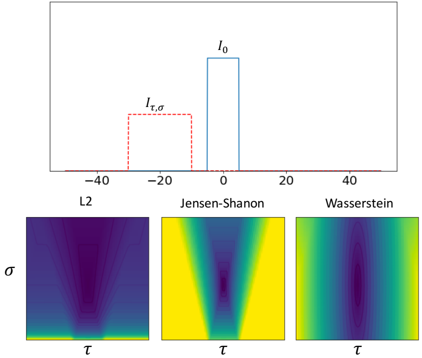
  or equivalently when the optimal transport map, f^*, exists,

$$W_p(\mu,\nu) = \left(\min_{f \in MP} \int_{\Omega} d^p(x, f(x)) d\mu(x) \right)^{\frac{1}{p}}.$$

- In most engineering applications $\Omega \subset \mathbb{R}^d$ and $d(x,y) = |x-y|$.

Why p-Wasserstein distance?





Optimality and Characterization (1D)

Theorem (optimality, uniqueness, monotonicity)

Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and suppose that corresponding (KP) is finite. Then (KP) has a unique solution γ^* characterized by the following monotonicity property:

$$(x, y), (x', y') \in \operatorname{supp}(\gamma), x < x' \Rightarrow y < y'. \tag{2}$$

Optimality and Characterization (1D)

Theorem (optimality, uniqueness, monotonicity)

Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and suppose that corresponding (KP) is finite. Then (KP) has a unique solution γ^* characterized by the following monotonicity property:

$$(x, y), (x', y') \in \operatorname{supp}(\gamma), x < x' \Rightarrow y < y'. \tag{2}$$

Moreover, if μ is atomless (does not give mass to atoms), this optimal plan is induced by a unique non-decreasing map T^*, i.e., $\gamma^* = (Id, T^*)_\sharp \mu$, in which case T^* is the minimizer (optimal transport map) for (MP).

Optimality and Characterization (1D)

Theorem (optimality, uniqueness, monotonicity)

Given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and suppose that corresponding (KP) is finite. Then (KP) has a unique solution γ^* characterized by the following monotonicity property:

$$(x, y), (x', y') \in \operatorname{supp}(\gamma), x < x' \Rightarrow y < y'. \quad (2)$$

Moreover, if μ is atomless (does not give mass to atoms), this optimal plan is induced by a unique non-decreasing map T^*, i.e., $\gamma^* = (Id, T^*)_\sharp \mu$, in which case T^* is the minimizer (optimal transport map) for (MP).

Remark:

1. See Section 1.6, 2.1 and 2.2 in F.Santambrogio's book *Optimal Transport for Applied Mathematicians.*
2. A key fact for proving the above theorem is that the support of an optimal plan is cyclically monotone, which can be then used to show (2).

Computation of Optimal Transport Map

Theorem

Given $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and suppose that μ is atomless. Then, the optimal transport map between them is the unique non-decreasing map $T^* : \mathbb{R} \to \mathbb{R}$ defined by

$$T^*(x) := F_\nu^\dagger(F_\mu(x)), \tag{3}$$

where F_μ denotes the cumulative distribution function of μ and $F_\nu^\dagger$ is the generalized/pseudo inverse function of F_ν.[1]

[1] The generalized inverse for a function $F : \mathbb{R} \to [0,1]$ is defined by $F^\dagger(x) := \inf\{t \in \mathbb{R} : F(t) \geq x\}$.

Computation of Optimal Transport Map

Theorem

Given $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and suppose that μ is atomless. Then, the optimal transport map between them is the unique non-decreasing map $T^* : \mathbb{R} \to \mathbb{R}$ defined by

$$T^*(x) := F_\nu^\dagger(F_\mu(x)), \tag{3}$$

where F_μ denotes the cumulative distribution function of μ and $F_\nu^\dagger$ is the generalized/pseudo inverse function of F_ν.[1]

Given L^1-normalized non-negative functions $f_\mu, f_\nu \in L^1(\mathbb{R})$, one can think of them as density functions of probability measures and define the corresponding concepts: (KP), (MP), optimal transport maps etc.

[1] The generalized inverse for a function $F : \mathbb{R} \to [0, 1]$ is defined by $F^\dagger(x) := \inf\{t \in \mathbb{R} : F(t) \geq x\}$.

p-Wasserstein distance for 1D probability measures

$$W_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt\right)^{\frac{1}{p}} \tag{4}$$

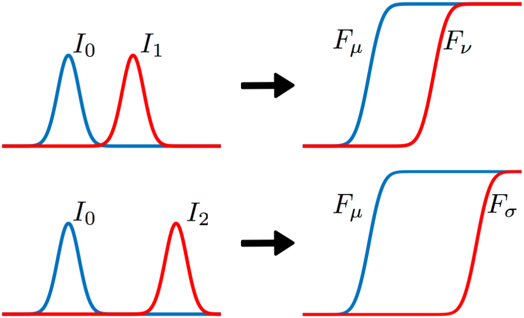


Figure: Note that, the Euclidean distance does not provide a sensible distance between I_0, I_1 and I_2 while the p-Wasserstein distance does.

p-Wasserstein distance for 1D probability measures

$$W_p(\mu,\nu) = (\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt)^{\frac{1}{p}} \quad (5)$$

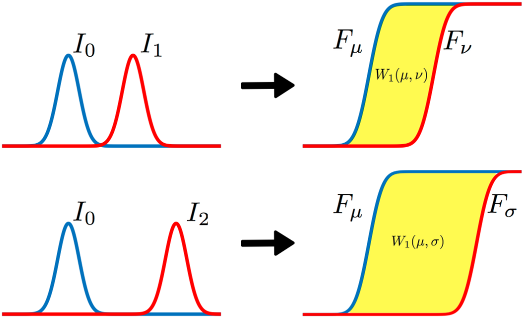
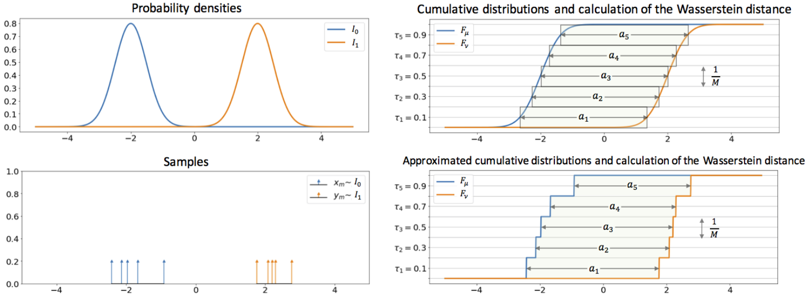


Figure: Note that, the Euclidean distance does not provide a sensible distance between I_0, I_1 and I_2 while the p-Wasserstein distance does.

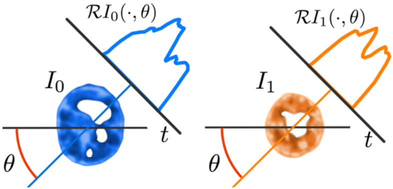
p-Wasserstein distance for 1D probability measures

$$W_p(\mu,\nu) = \left(\frac{1}{M}\sum_{m=1}^{M}\underbrace{(|F_\mu^{-1}(\tau_m) - F_\nu^{-1}(\tau_m)|)^p}_{a_m}\right)^{\frac{1}{p}} \tag{6}$$



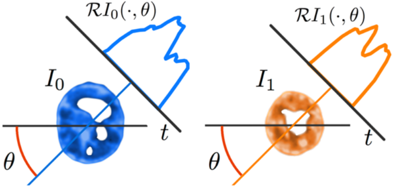
Sliced p-Wasserstein distance

- Slice an n-dimensional probability distribution $(n > 1)$ into one-dimensional representations through projections and measure p-Wasserstein distance between these representations.



Sliced p-Wasserstein distance

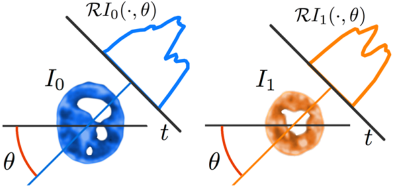
- Slice an n-dimensional probability distribution $(n > 1)$ into one-dimensional representations through projections and measure p-Wasserstein distance between these representations.



  - Where $\mathscr{R}$ denotes Radon transform and is defined as,

$$\mathscr{R}I(t,\theta) \quad := \quad \int_{\mathbb{S}^{d-1}} I(x)\delta(t-\theta\cdot x)dx$$
$$\forall t \in \mathbb{R},\ \forall\theta \in \mathbb{S}^{d-1}(\text{Unit sphere in } \mathbb{R}^d) \tag{7}$$

Sliced p-Wasserstein distance



- Slice an n-dimensional probability distribution ($n > 1$) into one-dimensional representations through projections and measure p-Wasserstein distance between these representations.
  - Where $\mathscr{R}$ denotes Radon transform and is defined as,

$$\mathscr{R}I(t,\theta) \quad := \quad \int_{\mathbb{S}^{d-1}} I(x)\delta(t-\theta\cdot x)dx$$
$$\forall t \in \mathbb{R}, \ \forall \theta \in \mathbb{S}^{d-1} (\text{Unit sphere in } \mathbb{R}^d) \tag{7}$$

  - and the p-Sliced-Wasserstein (p-SW) distance is defined as:

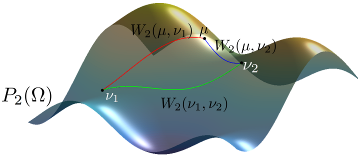
$$SW_p(I_0,I_1) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mathscr{R}I_0(.,\theta),\mathscr{R}I_1(.,\theta))d\theta\right)^{\frac{1}{p}} \tag{8}$$

Geometric Properties

2-Wasserstein geodesics

- The set of continuous measures together with the 2-Wasserstein metric forms a Riemmanian manifold.

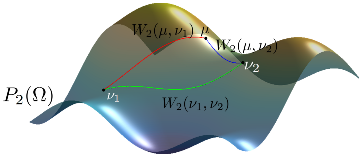
- Given the 2-Wasserstein space, $(P_2(\Omega), W_2)$, the geodesic between $\mu, \nu \in P_2(\Omega)$ is the shortest curve on $P_2(\Omega)$ that connects these measures.



2-Wasserstein geodesics

- The set of continuous measures together with the 2-Wasserstein metric forms a Riemmanian manifold.

- Given the 2-Wasserstein space, $(P_2(\Omega), W_2)$, the geodesic between $\mu, \nu \in P_2(\Omega)$ is the shortest curve on $P_2(\Omega)$ that connects these measures.

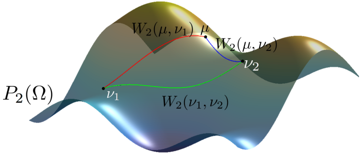


- Let ρ_t for $t \in [0,1]$ parametrizes a curve on $P_2(\Omega)$ with $\rho_0 = \mu$ and $\rho_1 = \nu$, and let I_t denote the density of ρ_t, $I_t(x)dx = d\rho_t(x)$.

2-Wasserstein geodesics

- The set of continuous measures together with the 2-Wasserstein metric forms a Riemmanian manifold.

- Given the 2-Wasserstein space, $(P_2(\Omega), W_2)$, the geodesic between $\mu, \nu \in P_2(\Omega)$ is the shortest curve on $P_2(\Omega)$ that connects these measures.



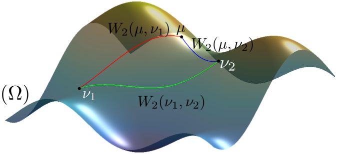
- Let ρ_t for $t \in [0, 1]$ parametrizes a curve on $P_2(\Omega)$ with $\rho_0 = \mu$ and $\rho_1 = \nu$, and let I_t denote the density of ρ_t, $I_t(x)dx = d\rho_t(x)$.
- For the optimal transport map, $f(x)$, between μ and ν the geodesic is parametrized as,

$$I_t(x) = det(Df_t(x))I_1(f_t(x)), \quad f_t(x) = (1-t)x + tf(x) \tag{9}$$

2-Wasserstein geodesics

- The set of continuous measures together with the 2-Wasserstein metric forms a Riemmanian manifold.

- Given the 2-Wasserstein space, $(P_2(\Omega), W_2)$, the geodesic between $\mu, \nu \in P_2(\Omega)$ is the shortest curve on $P_2(\Omega)$ that connects these measures.



- Let ρ_t for $t \in [0,1]$ parametrizes a curve on $P_2(\Omega)$ with $\rho_0 = \mu$ and $\rho_1 = \nu$, and let I_t denote the density of ρ_t, $I_t(x)dx = d\rho_t(x)$.
- For the optimal transport map, $f(x)$, between μ and ν the geodesic is parametrized as,

$$I_t(x) \quad = \quad det(Df_t(x))I_1(f_t(x)), \quad f_t(x) = (1-t)x + tf(x) \tag{9}$$

- It is straightforward to show that,

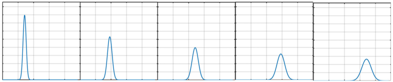
$$W_2(\mu, \rho_t) = tW_2(\mu, \nu) \tag{10}$$

2-Wasserstein geodesics

Geodesic in the 2-Wasserstein space

$$I_t(x) = det(Df_t(x))I_1(f_t(x))$$

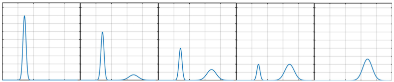
$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$



Geodesic in the Euclidean space

$$I_t(x) = (1-t)I_0(x) + tI_1(x)$$

$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$

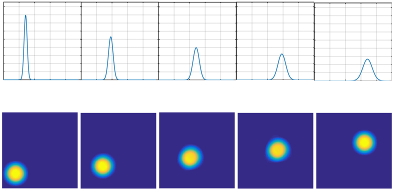


2-Wasserstein geodesics

Geodesic in the 2-Wasserstein space

$$I_t(x) = det(Df_t(x))I_1(f_t(x))$$

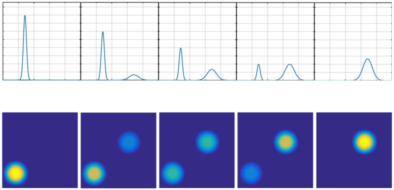
$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$



Geodesic in the Euclidean space

$$I_t(x) = (1 - t)I_0(x) + tI_1(x)$$

$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$

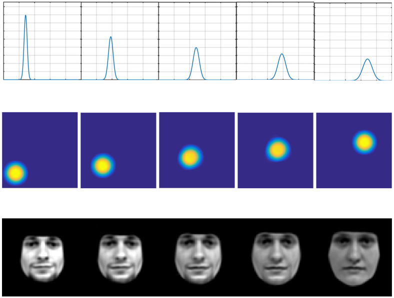


2-Wasserstein geodesics

Geodesic in the 2-Wasserstein space

$$I_t(x) = det(Df_t(x))I_1(f_t(x))$$

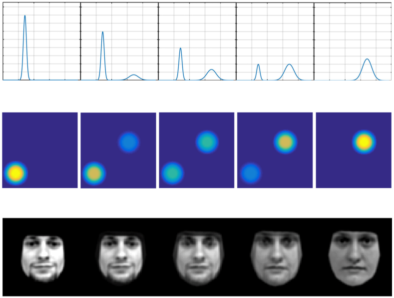
$t=0$ $t=0.25$ $t=0.5$ $t=0.75$ $t=1$



Geodesic in the Euclidean space

$$I_t(x) = (1-t)I_0(x) + tI_1(x)$$

$t=0$ $t=0.25$ $t=0.5$ $t=0.75$ $t=1$

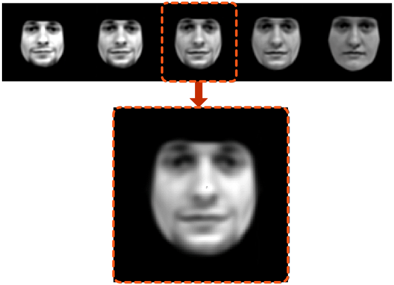


2-Wasserstein geodesics

Geodesic in the 2-Wasserstein space

$$I_t(x) = det(Df_t(x))I_1(f_t(x))$$

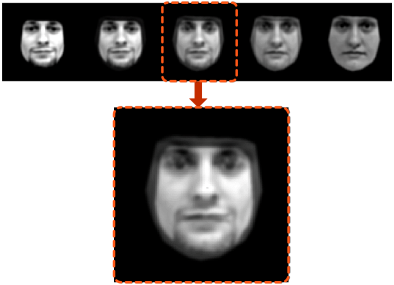
$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$



Geodesic in the Euclidean space

$$I_t(x) = (1-t)I_0(x) + tI_1(x)$$

$t = 0$ $t = 0.25$ $t = 0.5$ $t = 0.75$ $t = 1$



Numerical Solvers

Flow Minimization (Angenent, Haker, and Tannenbaum)

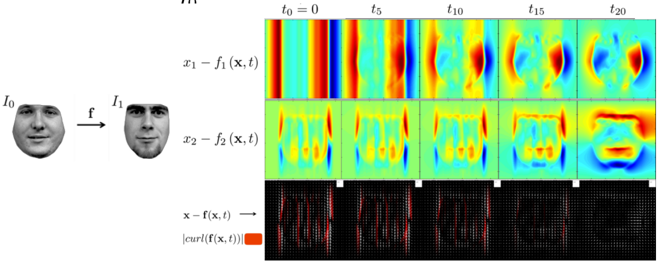
- The flow minimization method finds the optimal transport map following below steps:
  1. Obtain an initial mass preserving transport map using the Knothe-Rosenblatt coupling
  2. Update the initial map to obtain a curl free mass preserving transport map that minimizes the transport cost

$$f_{k+1} = f_k + \epsilon \frac{1}{I_0} D f_k (f_k - \nabla(\Delta^{-1} div(f_k))), \quad \Delta^{-1} : \text{Poisson solver} \quad (11)$$

Flow Minimization (Angenent, Haker, and Tannenbaum)

- The flow minimization method finds the optimal transport map following below steps:
    1. Obtain an initial mass preserving transport map using the Knothe-Rosenblatt coupling
    2. Update the initial map to obtain a curl free mass preserving transport map that minimizes the transport cost

$$f_{k+1} = f_k + \epsilon \frac{1}{I_0} D f_k (f_k - \nabla(\Delta^{-1} div(f_k))), \quad \Delta^{-1} : \text{Poisson solver} \quad (11)$$



Angenent, S., et al. "Minimizing flows for the Monge–Kantorovich problem." SIAM 2003

Gradient descent on the dual problem (Chartrand et al.)

- For the strictly convex cost function, $c(x,y) = \frac{1}{2}|x-y|^2$, the dual of Kantorovich problem can be formalized as minimizing,

$$M(\eta) = \int_X \phi(x) d\mu(x) + \int_Y \phi^c(y) d\nu(y) \tag{12}$$

$\eta^c(y) := \max_{x \in X}(x \cdot y - \phi(x))$ is the Legendre-Fenchel transform of $\phi(x)$.

Gradient descent on the dual problem (Chartrand et al.)

- For the strictly convex cost function, $c(x,y) = \frac{1}{2}|x-y|^2$, the dual of Kantorovich problem can be formalized as minimizing,

$$M(\eta) = \int_X \phi(x)d\mu(x) + \int_Y \phi^c(y)d\nu(y) \tag{12}$$

$\eta^c(y) := \max_{x\in X}(x \cdot y - \phi(x))$ is the Legendre-Fenchel transform of $\phi(x)$.

- Then the potential transport field, ϕ, is updated to minimize $M(\phi)$ through,

$$\phi_{k+1} = \phi_k - \epsilon(I_0 - \det(I - H\phi_k^{cc})I_1(id - \nabla\phi_k^{cc})), \quad H\text{: Hessian matrix} \tag{13}$$

Gradient descent on the dual problem (Chartrand et al.)

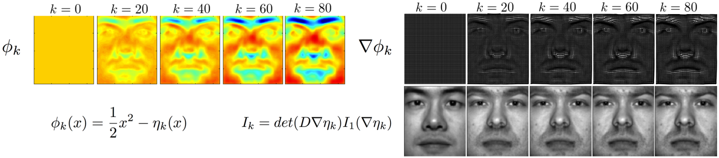
- For the strictly convex cost function, $c(x,y) = \frac{1}{2}|x-y|^2$, the dual of Kantorovich problem can be formalized as minimizing,

$$M(\eta) = \int_X \phi(x) d\mu(x) + \int_Y \phi^c(y) d\nu(y) \tag{12}$$

$\eta^c(y) := \max_{x \in X}(x \cdot y - \phi(x))$ is the Legendre-Fenchel transform of $\phi(x)$.

- Then the potential transport field, ϕ, is updated to minimize $M(\phi)$ through,

$$\phi_{k+1} = \phi_k - \epsilon(I_0 - \det(I - H\phi_k^{cc}) I_1(id - \nabla\phi_k^{cc})), \quad H\text{: Hessian matrix} \tag{13}$$

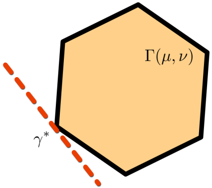


Chartrand, R., et al. "A gradient descent solution to the Monge-Kantorovich problem." AMS 2009

Linear programming

- Let $\mu = \sum_{i=1}^{N} p_i \delta_{x_i}$ and $\mu = \sum_{j=1}^{M} q_j \delta_{y_j}$, where δ_{x_i} is a Dirac measure,

$$KP(\mu,\nu) = \min_\gamma \sum_i \sum_j c(x_i, y_j)\gamma_{ij}$$
$$s.t. \quad \sum_j \gamma_{ij} = p_i,\ \sum_i \gamma_{ij} = q_j,\ \gamma_{ij} \geq 0$$

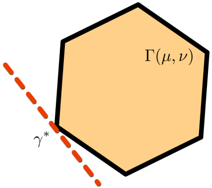


  - Can be solved through the Simplex algorithm or interior point techniques.
  - Computational complexity of solvers: $\mathcal{O}(N^3 logN)$

Linear programming

- Let $\mu = \sum_{i=1}^{N} p_i \delta_{x_i}$ and $\mu = \sum_{j=1}^{M} q_j \delta_{y_j}$, where δ_{x_i} is a Dirac measure,

$$KP(\mu,\nu) \quad = \quad \min_\gamma \sum_i \sum_j c(x_i, y_j)\gamma_{ij}$$
$$s.t. \quad \sum_j \gamma_{ij} = p_i, \ \sum_i \gamma_{ij} = q_j, \ \gamma_{ij} \geq 0$$



  - Can be solved through the Simplex algorithm or interior point techniques.
  - Computational complexity of solvers: $\mathcal{O}(N^3 log N)$

Multi-Scale Approaches

- To improve computational complexity of several multi-scale approaches have been proposed
- The idea behind all these multi-scale techniques is to obtain a coarse transport plan and refine the transport plan iteratively.

Entropy Regularization

- Cuturi proposed a regularized version of the Kantorovich problem which can be solved in $\mathcal{O}(NlogN)$,

$$W_{p,\lambda}^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega\times\Omega} d^p(x,y)\gamma(x,y) + \lambda\gamma(x,y)\ln(\gamma(x,y))dxdy. \quad (14)$$

Entropy Regularization

- Cuturi proposed a regularized version of the Kantorovich problem which can be solved in $\mathcal{O}(NlogN)$,

$$W_{p,\lambda}^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega\times\Omega} d^p(x,y)\gamma(x,y) + \lambda\gamma(x,y)\ln(\gamma(x,y))dxdy. \quad (14)$$

- It is straightforward to show that the entropy regularized p-Wasserstein distance in Equation (14) can be reformulated as,

$$W_{p,\lambda}^p(\mu,\nu) = \lambda \inf_{\gamma\in\Gamma(\mu,\nu)} \mathsf{KL}(\gamma|\mathcal{K}_\lambda), \qquad \mathcal{K}_\lambda(x,y) = \exp(-\frac{d^p(x,y)}{\lambda}) \quad (15)$$

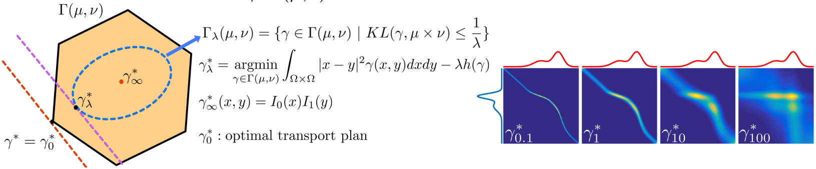
Entropy Regularization

- Cuturi proposed a regularized version of the Kantorovich problem which can be solved in $\mathcal{O}(NlogN)$,

$$W_{p,\lambda}^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\Omega\times\Omega} d^p(x,y)\gamma(x,y) + \lambda\gamma(x,y)\ln(\gamma(x,y))dxdy. \quad (14)$$

- It is straightforward to show that the entropy regularized p-Wasserstein distance in Equation (14) can be reformulated as,

$$W_{p,\lambda}^p(\mu,\nu) = \lambda \inf_{\gamma\in\Gamma(\mu,\nu)} \mathrm{KL}(\gamma|\mathcal{K}_\lambda), \qquad \mathcal{K}_\lambda(x,y) = \exp(-\frac{d^p(x,y)}{\lambda}) \quad (15)$$



Cuturi, M. "Sinkhorn distances: Lightspeed computation of optimal transport." NIPS 2013.

Summary

- Introduction:
  1. Monge formulation $\rightarrow$ transport maps
  2. Kantorovich formulation $\rightarrow$ transport plans

Summary

- Introduction:
  1. Monge formulation $\rightarrow$ transport maps
  2. Kantorovich formulation $\rightarrow$ transport plans
- Transport-based metrics and geometric properties:
  1. p-Wasserstein metric
  2. p-Sliced Wasserstein metric
  3. 2-Wasserstein geodesics

Summary

- Introduction:
  1. Monge formulation $\rightarrow$ transport maps
  2. Kantorovich formulation $\rightarrow$ transport plans
- Transport-based metrics and geometric properties:
  1. p-Wasserstein metric
  2. p-Sliced Wasserstein metric
  3. 2-Wasserstein geodesics
- Numerical solvers:
  1. Flow minimization (Monge)
  2. Gradient descent on the dual problem (Monge)
  3. Linear programming and multi-scale methods (Kantorovich)
  4. Entropy regularized solver (Kantorovich)

Summary

- Introduction:
  1. Monge formulation $\rightarrow$ transport maps
  2. Kantorovich formulation $\rightarrow$ transport plans
- Transport-based metrics and geometric properties:
  1. p-Wasserstein metric
  2. p-Sliced Wasserstein metric
  3. 2-Wasserstein geodesics
- Numerical solvers:
  1. Flow minimization (Monge)
  2. Gradient descent on the dual problem (Monge)
  3. Linear programming and multi-scale methods (Kantorovich)
  4. Entropy regularized solver (Kantorovich)

Coming Up Next

- Transport-based transformations

Thank you!