Introduction to Optimal Transport

Larry Wasserman
Department of Statistics and Data Science
Carnegie Mellon
larry@stat.cmu.edu

Introduction: What is Optimal Transport?

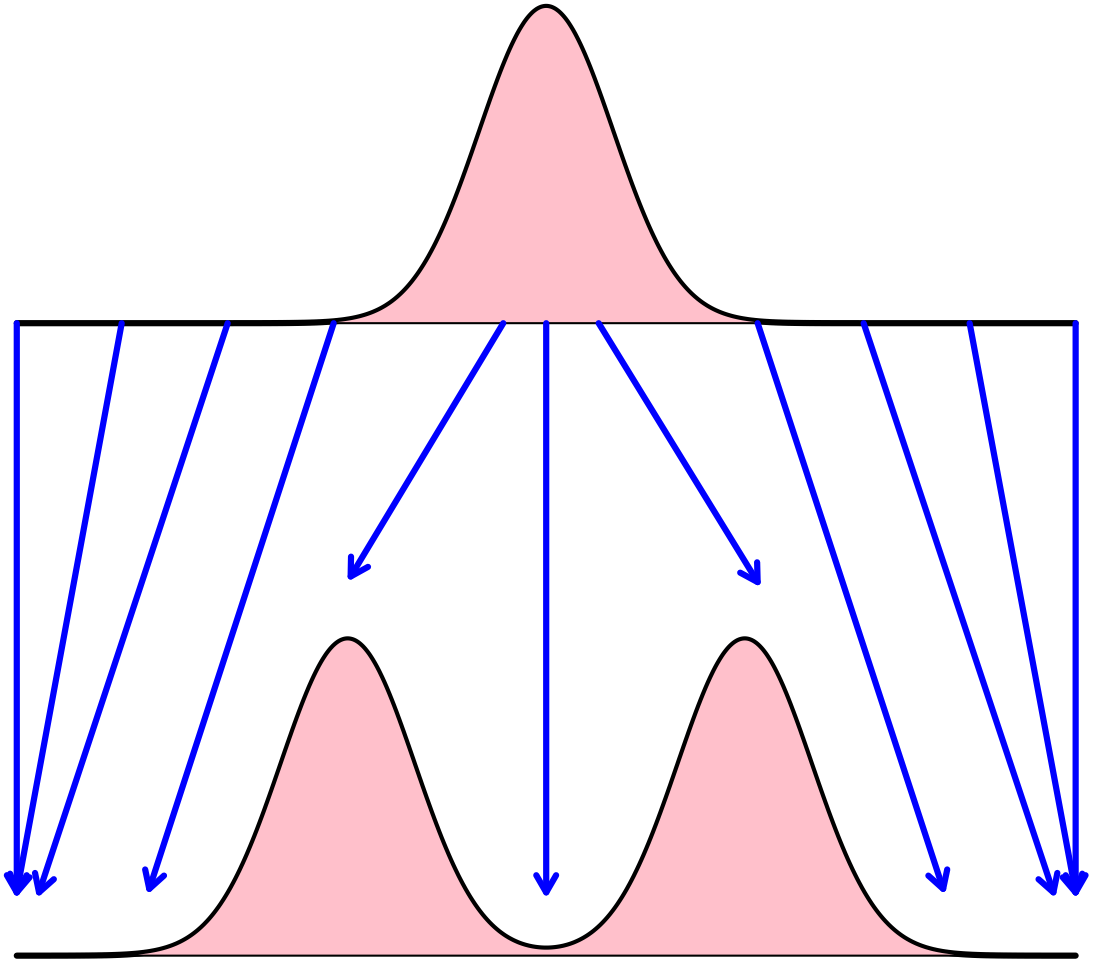
We have two distributions P_0 and P_1.

Introduction: What is Optimal Transport?

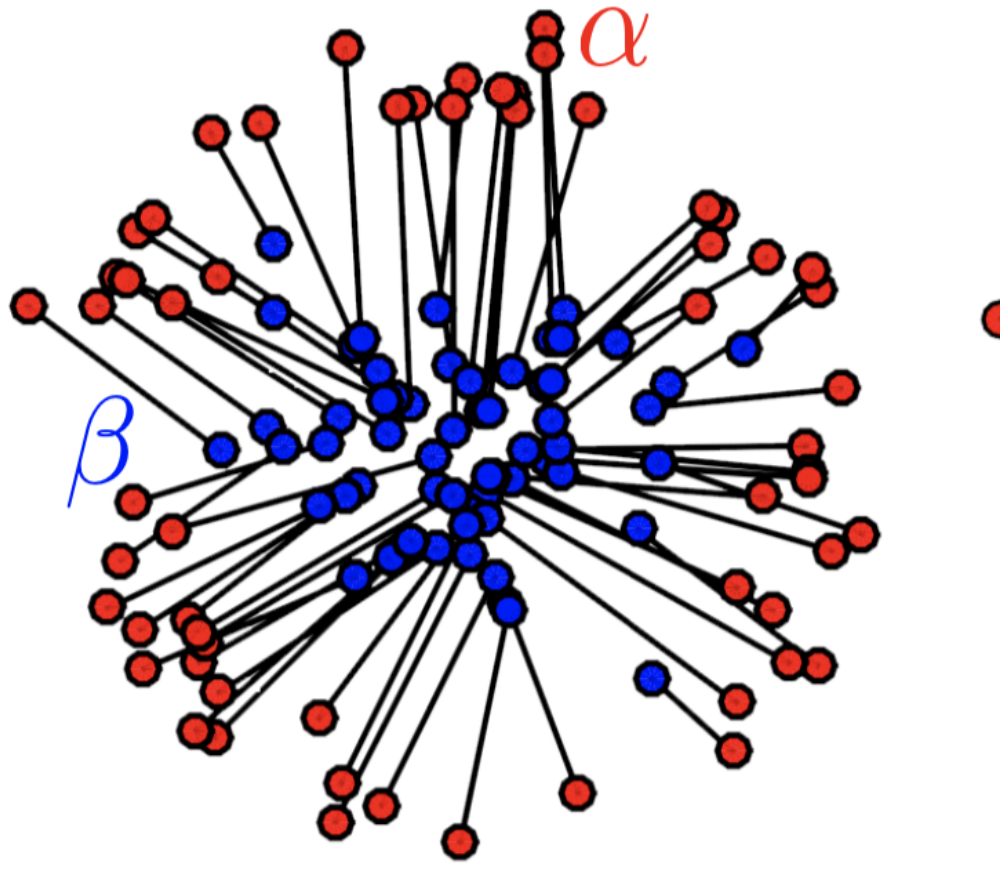
We have two distributions P_0 and P_1.
Goals:

- Define an "optimal map" that transforms P_0 into P_1.
- Define a distance based on transport (Wasserstein distance)
- Define a path (geodesic) between P_1 and P_2 (morphing) in the space of distributions.
- Define a shape-preserving notion of" averages" of distributions.

Optimal Transport (Monge 1781)



Point Cloud Example (from Peyre, Cuturi 2019)



Optimal Transport: Monge Version

Let $X \sim P_0$.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance.

For now, assume that the minimizer exists. The the minimizer is called the optimal transport map.

Optimal Transport: Monge Version

Let $X \sim P_0$.
Find T to minimize

$$\mathbb{E}\left[||X - T(X)||^p\right] = \int ||x - T(x)||^p dP_0(x)$$

over all maps T such that $T(X) \sim P_1$.
Can replace Euclidean distance with any distance.

For now, assume that the minimizer exists. The the minimizer is called the optimal transport map.

Common choices: $p = 2$ or $p = 1$.

Wasserstein (transport) distance

$$W_p(X, Y) \equiv W_p(P_0, P_1) = \left(\int ||x - T^*(x)||^p dP_0(x) \right)^{1/p}$$

where T^* is the optimal transport map.

Defines a metric on the space of (nearly) all distributions.

W_1 is called the Earth Mover Distance

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$

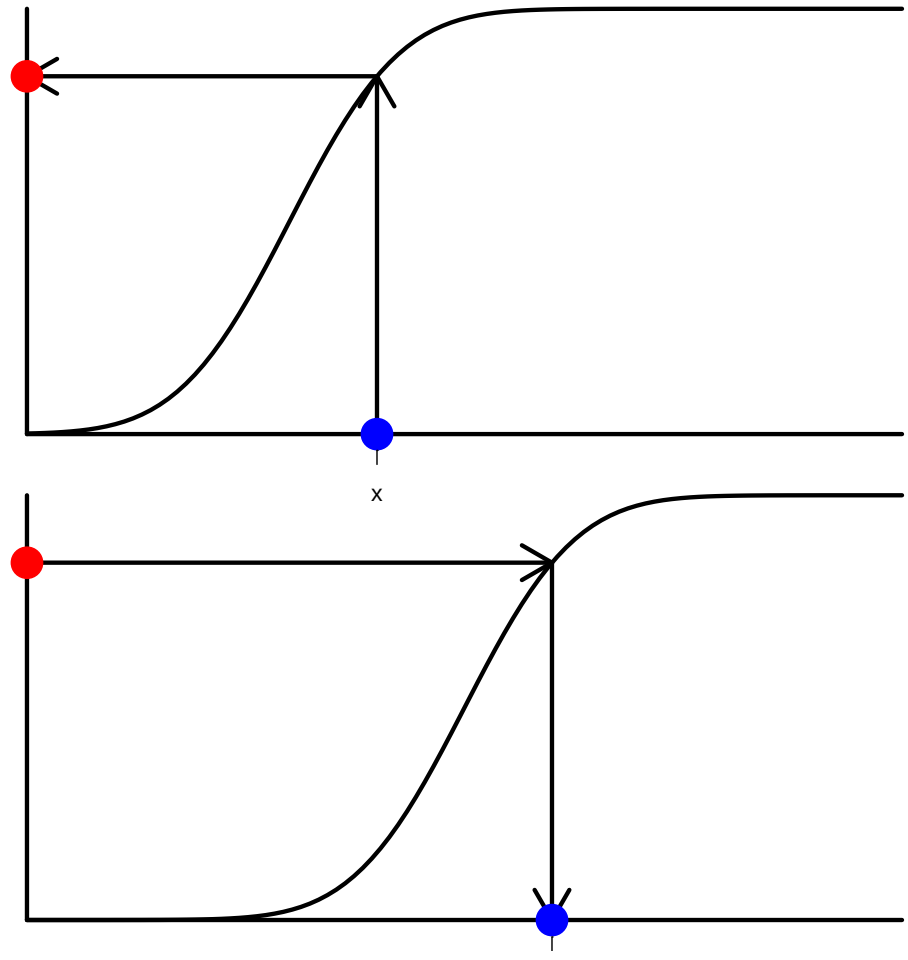
Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$
- $W_p(P_0, P_1) = \left(\int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds\right)^{1/p}$

Finding the Transport Map: One Dimensional Case

- Find the cdf (cumulative distribution function)
- $F_0(s) = P_0(X \leq s)$
- $F_1(s) = P_1(Y \leq s)$
- The optimal map is: $T(s) = F_1^{-1}\big(F_0(s)\big)$
- $W_p(P_0, P_1) = \left(\int |F_0^{-1}(s) - F_1^{-1}(s)|^p ds\right)^{1/p}$
- The morphing — geodesic linking F_0 and F_1 — is

$$F_s = \left[(1-s)F_0^{-1} + sF_1^{-1}\right]^{-1}$$


x

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

$$\widehat{F}_0(s) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq s)$$

Data Version

$X_1, \ldots, X_n \sim P_0$
$Y_1, \ldots, Y_m \sim P_1$

Just substitute the estimated (empirical) cdf's:

$$\widehat{F}_0(s) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq s)$$

$$\widehat{F}_1(s) = \frac{1}{m} \sum_{i=1}^{m} I(Y_i \leq s)$$

Finding the Transport Map: Gaussian Case

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Then:

$$T(X) = \mu_1 + \Sigma_1^{1/2}\Sigma_0^{-1/2}(X - \mu_0)$$

Finding the Transport Map: Gaussian Case

If $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

Then:

$$T(X) = \mu_1 + \Sigma_1^{1/2}\Sigma_0^{-1/2}(X - \mu_0)$$

$$W_2^2(P_0, P_1) = ||\mu_0 - \mu_1||^2 + B(\Sigma_0, \Sigma_1)$$

where

$$B(\Sigma_0, \Sigma_1) = \text{trace}(\Sigma_0) + \text{trace}(\Sigma_1) - 2\text{trace}\big[(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\big].$$

Finding the Transport Map: Two Point Clouds

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$
- $\mathcal{Y} = \{Y_1, \ldots, Y_n\} \quad Y_i \in \mathbb{R}^d$

Finding the Transport Map: Two Point Clouds

- $\mathcal{X} = \{X_1, \ldots, X_n\} \quad X_i \in \mathbb{R}^d$
- $\mathcal{Y} = \{Y_1, \ldots, Y_n\} \quad Y_i \in \mathbb{R}^d$
- $T : X_i \to Y_{\pi(i)}$ where π minimizes

$$\sum_i ||X_i - Y_{\pi(i)}||^2$$

  over all permutations π.
- Hungarian algorithm $O(n^3)$ time.

How Accurate is This?

$X_1, \ldots, X_n \sim P$

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

$\widehat{T}$ is estimated from data (and extended by one-nearest-neighbor):

How Accurate is This?

$X_1, \ldots, X_n \sim P$

$Y_1, \ldots, Y_n \sim Q$

T is true map from P to Q.

$\widehat{T}$ is estimated from data (and extended by one-nearest-neighbor):

under conditions (Manole, Balakrishnan and Wasserman, in progress):

$$\mathbb{E}||\widehat{T}(X) - T(X)||^2 = O(n^{-2/d})$$

and this is optimal without further conditions.

Smooth Transport

Smooth Transport

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.

Smooth Transport

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.

estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Smooth Transport

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.

estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

Smooth Transport

estimate p with kernel estimator $\widehat{p}_h$ using bandwidth h.

estimate q with kernel estimator $\widehat{q}_h$ using bandwidth h.

Sample from $\widehat{p}_h$ and $\widehat{q}_h$ and apply Hungarian algorithm.

Expensive but leverages the smoothness.

Other Computing Methods

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)
- Subsampling (Sommerfeld, Schrieber, Zemel and Munk, 2018)
- Hubs (Forrow et al 2018)

Other Computing Methods

- Sinkhorn (Cuturi 2013)
- Multiscale (Merigot 2011, Gerber and Maggioni 2017)
- Tangent space approximation (Wang, Slepcev, Basu, Ozolek, Rohde 2012)
- Slicing (Bonneel et al 2015)
- Subsampling (Sommerfeld, Schrieber, Zemel and Munk, 2018)
- Hubs (Forrow et al 2018)

- See: POT (Python Optimal Transport)

`https://pot.readthedocs.io/en/stable/`

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a geodesic between them.

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a geodesic between them.

- $T_s(x) = (1 - s)x + sT(x)$

Geodesics (Morphing)

- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a geodesic between them.

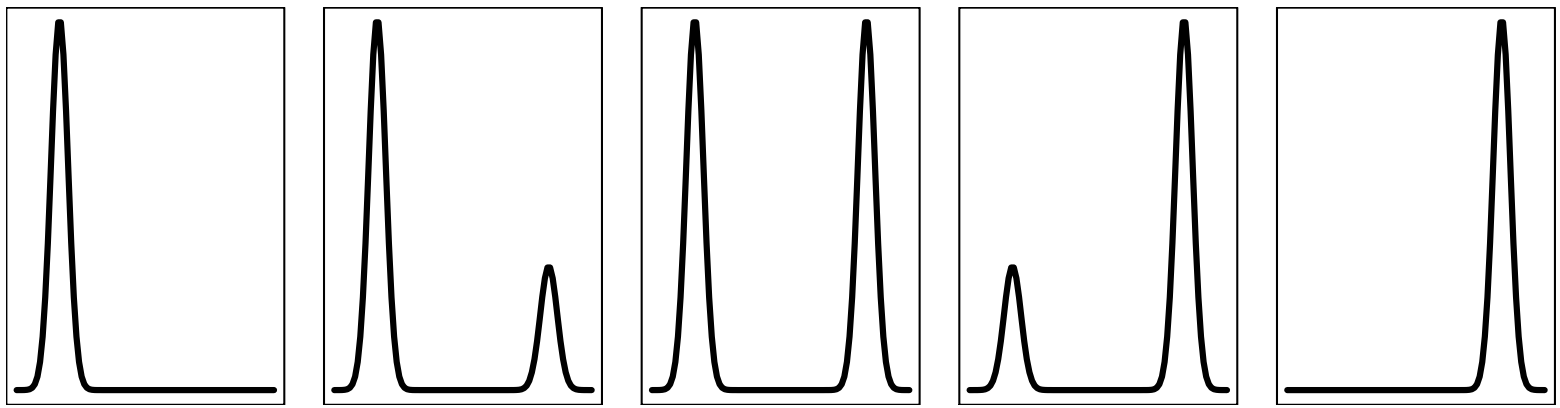
- $T_s(x) = (1 - s)x + sT(x)$
- $P_s = T_{s\#}P$.
  In other words, P_s is the distribution of the random variable $(1 - s)X + sT(X)$ where $X \sim P_0$.

Geodesics (Morphing)

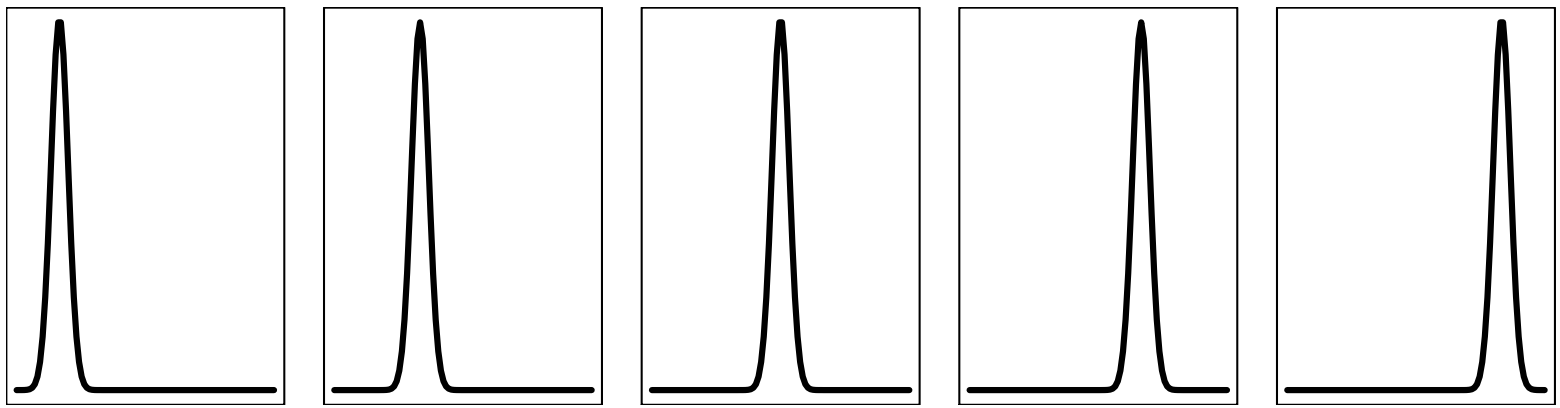
- The set of distributions $\mathcal{P}$ equipped with Wasserstein distance W is a geodesic space (and is Riemannian when $p = 2$).
- Given P_0 and P_1 there is a geodesic between them.

- $T_s(x) = (1 - s)x + sT(x)$
- $P_s = T_{s\#}P$.
  In other words, P_s is the distribution of the random variable $(1 - s)X + sT(X)$ where $X \sim P_0$.
- Then $(P_s : 0 \leq t \leq 1)$ is the geodesic.
  Length of the path $= W(P_0, P_1)$.

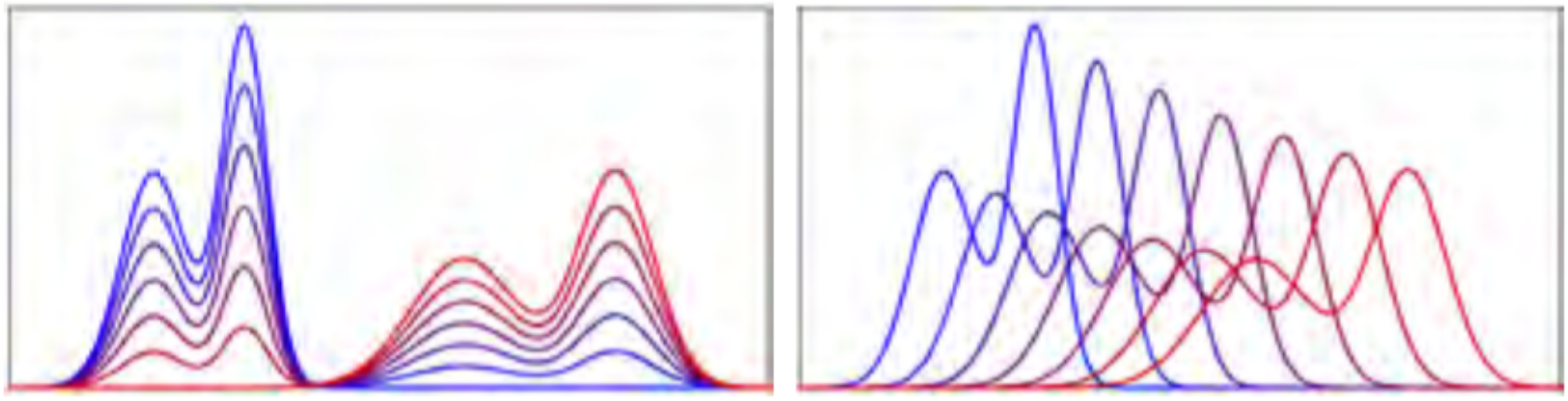
Euclidean Path between Two Gaussians



Geodesic Path between Two Gaussians



Geodesic Path between Two Mixtures: Bonneel, Peyre, Cuturi 2016



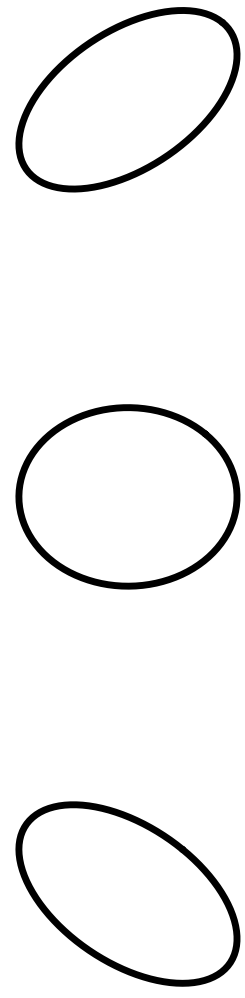
ℓ_2 interpolation

Wasserstein interpolation

Geodesic Path Between Two Images

Image credit: Bauer, Joshi and Modin 2015.

Bivariate Gaussian



Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Wasserstein barycenter: find P to minimize:

$$\sum_j W^2(P, P_j).$$

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

Same problem as before: this does not look like any of the P_j's.

Wasserstein barycenter: find P to minimize:

$$\sum_j W^2(P, P_j).$$

This is the barycenter and it is shape preserving.

Barycenters

Given $P_1, \ldots, P_N$, what is the 'average' of the P_j's?

Euclidean average?

$$\frac{1}{N} \sum_j P_j$$

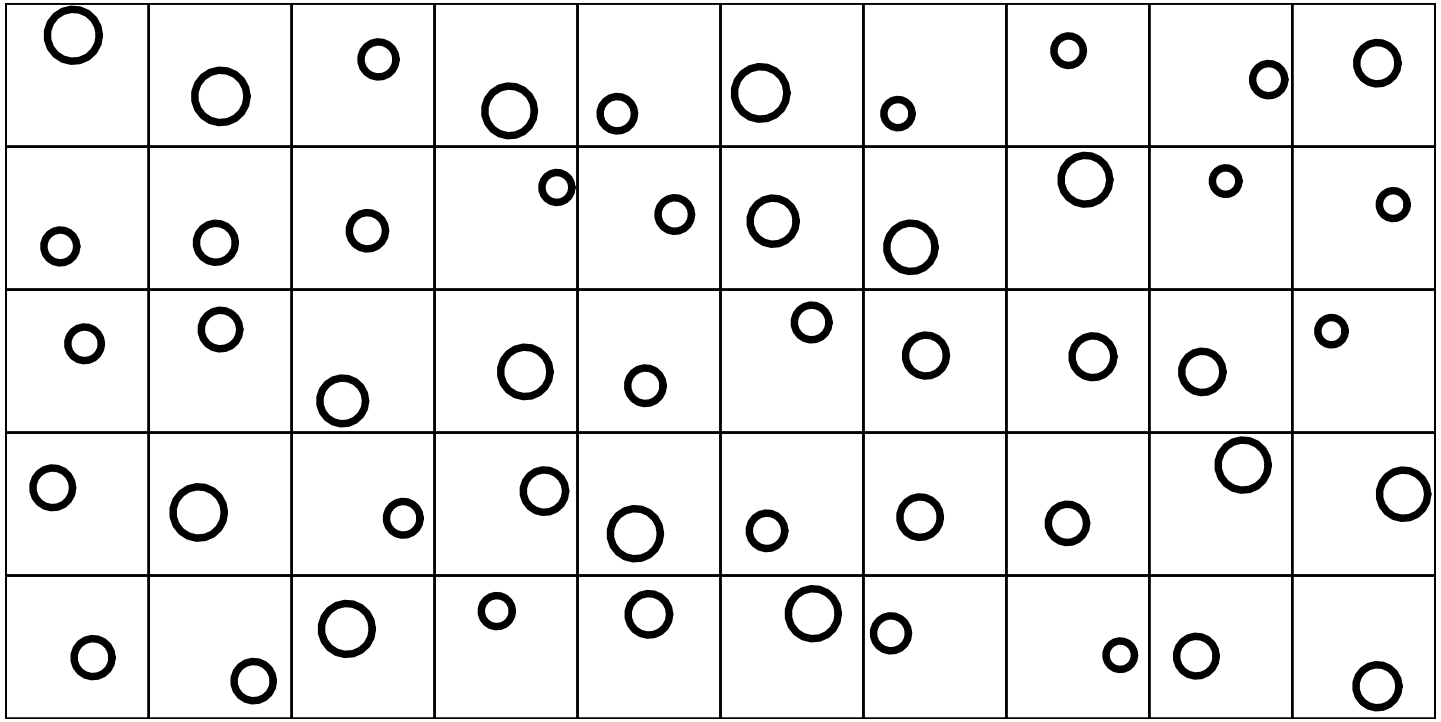
Same problem as before: this does not look like any of the P_j's.

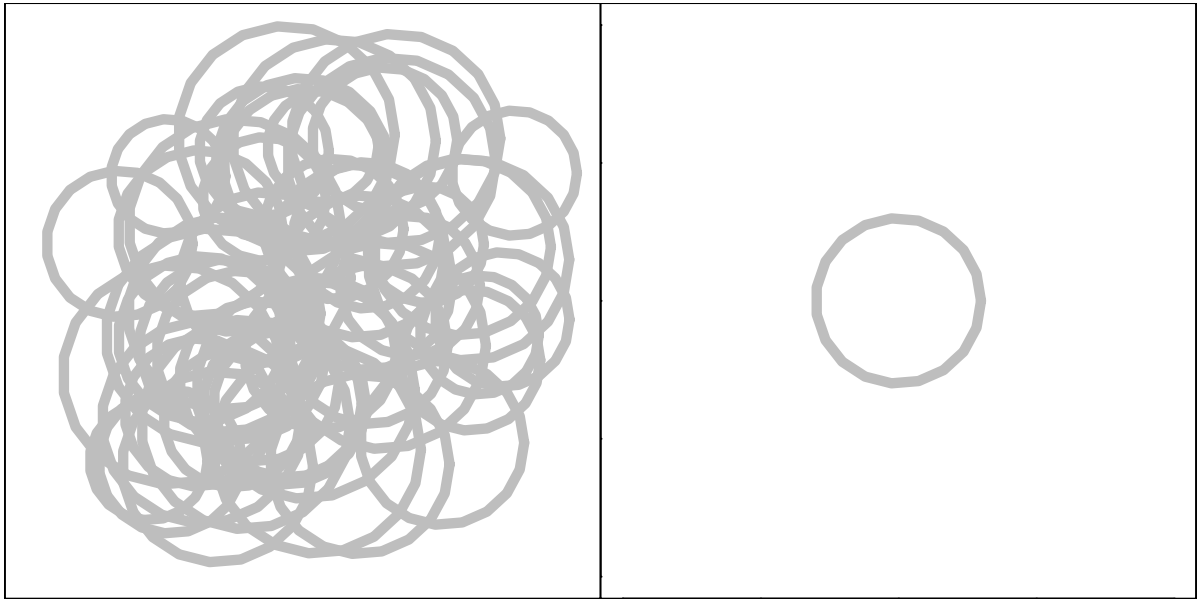
Wasserstein barycenter: find P to minimize:

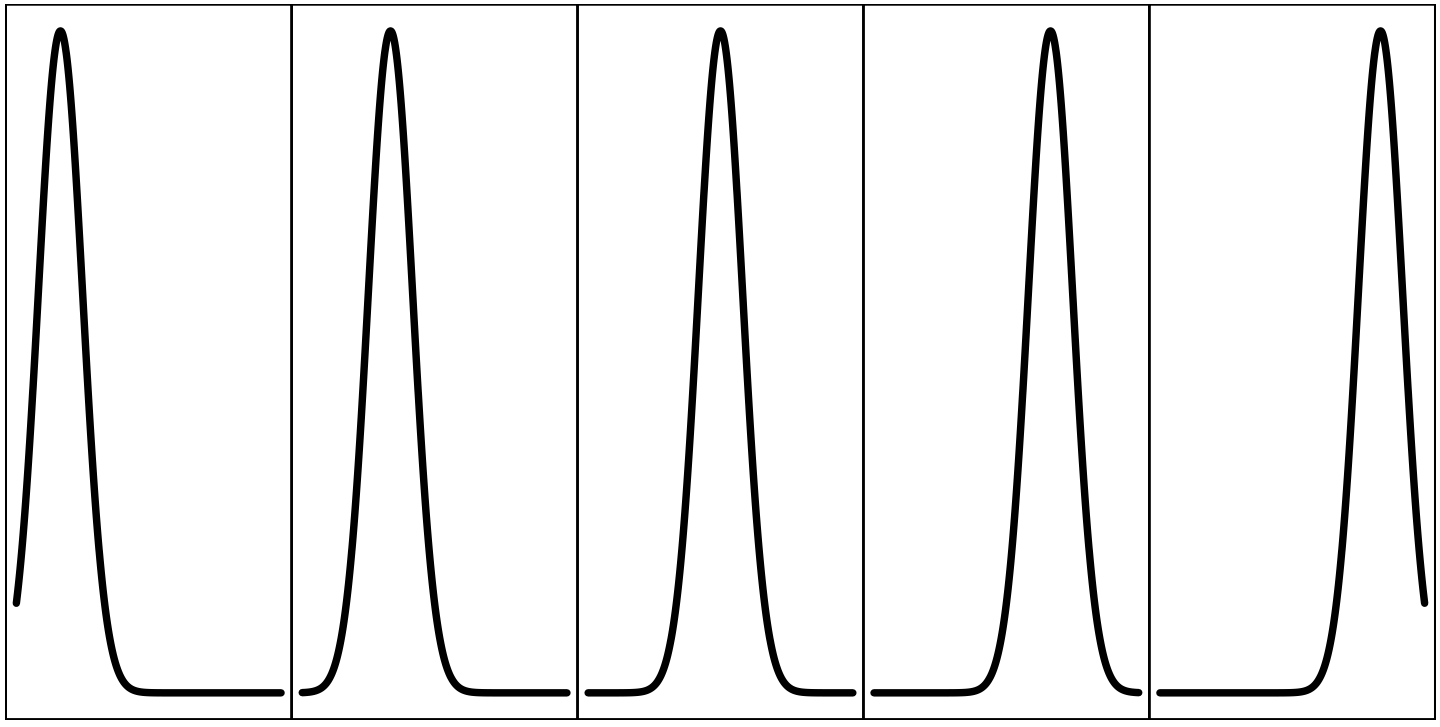
$$\sum_j W^2(P, P_j).$$

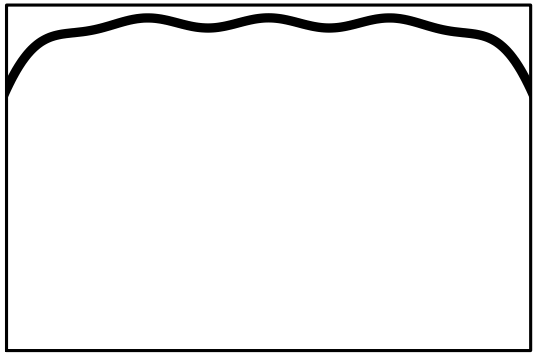
This is the barycenter and it is shape preserving.

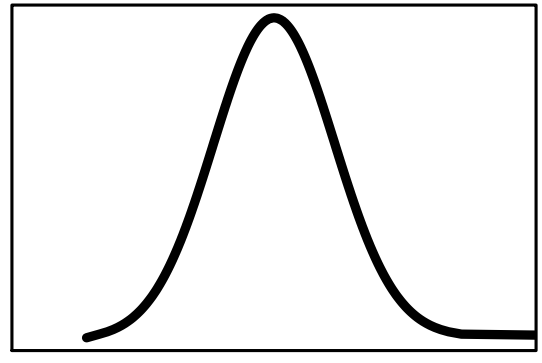
We can then define morphings from the barycenter to each of the P_j.



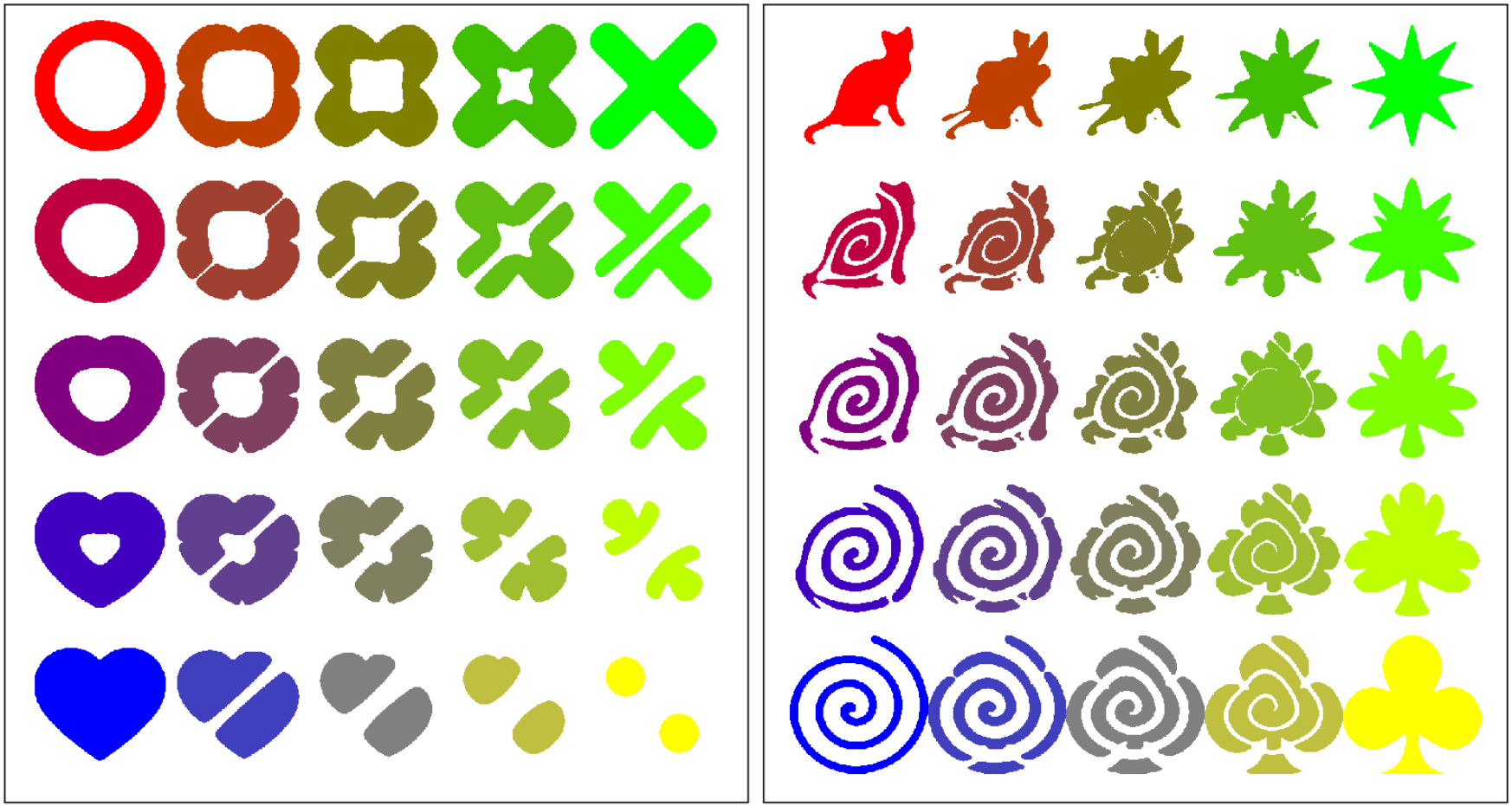








Example from Peyre and Cuturi 2019



How to Compute the Barycenter?

Active research area.

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

$\overline{F} = Q^{-1}$ where

$$Q(u) = \frac{1}{N} \sum_j F_j^{-1}(u)$$

How to Compute the Barycenter?

Active research area.

In one dimension it is easy:

$\overline{F} = Q^{-1}$ where

$$Q(u) = \frac{1}{N} \sum_j F_j^{-1}(u)$$

See Claici, Chien, Solomon (arXiv:1802.05757) and references therein.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Solution: Kantoravich relaxation:

Optimal Transport (Kantorovich Version, Transport Plans)

An important technical detail that we have ignored:

There may not be a map that takes P to Q.

For example, if $P = \delta_0$ (point mass at 0) and $Q =$Gaussian.

Solution: Kantoravich relaxation:

Take mass at x, and split it into many small pieces.

Optimal Transport (Kantorovich Version)

Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.

Optimal Transport (Kantorovich Version)

Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.
Find J (optimal transport plan) to minimize

$$\mathbb{E}_J[||X - Y||] = \left(\int ||x - y||^p \, dJ(x, y) \right)^{1/p}.$$

Optimal Transport (Kantorovich Version)

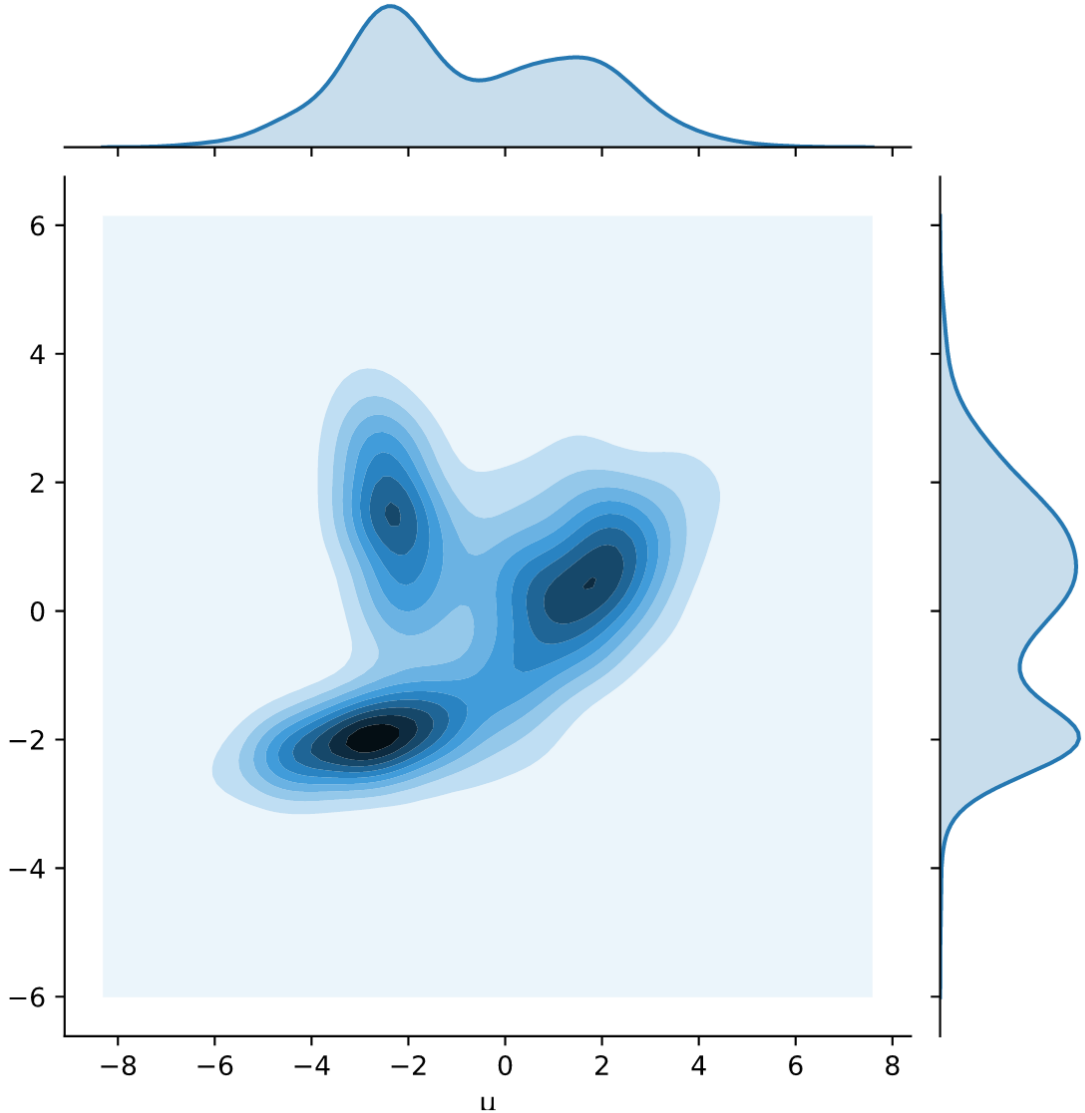
Let $\mathcal{J}$ denote all joint distributions J for (X, Y) with marginals P and Q. Each J is called a coupling between P and Q.
Find J (optimal transport plan) to minimize

$$\mathbb{E}_J[||X - Y||] = \left(\int ||x - y||^p \, dJ(x, y) \right)^{1/p}.$$

Again, this defines a distance

$$W(P, Q) = W(X, Y) = \left(\inf_J \int (||x - y||^2 dJ(x, y) \right)^{1/2}$$

called the Wasserstein distance.



Joint distribution J with a given X marginal and a given Y marginal. Image credit: Wikipedia.

Morphing

In this case, the morphing (geodesic) can be described as follows.

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

that is,

Morphing

In this case, the morphing (geodesic) can be described as follows.

Let J be the optimal transport plan for P_0 and P_1.

Let $F_s(x, y) = (1 - t)x + ty$

Then P_s is the distribution of $F_s(X, Y)$ where $(X, Y) \sim J$

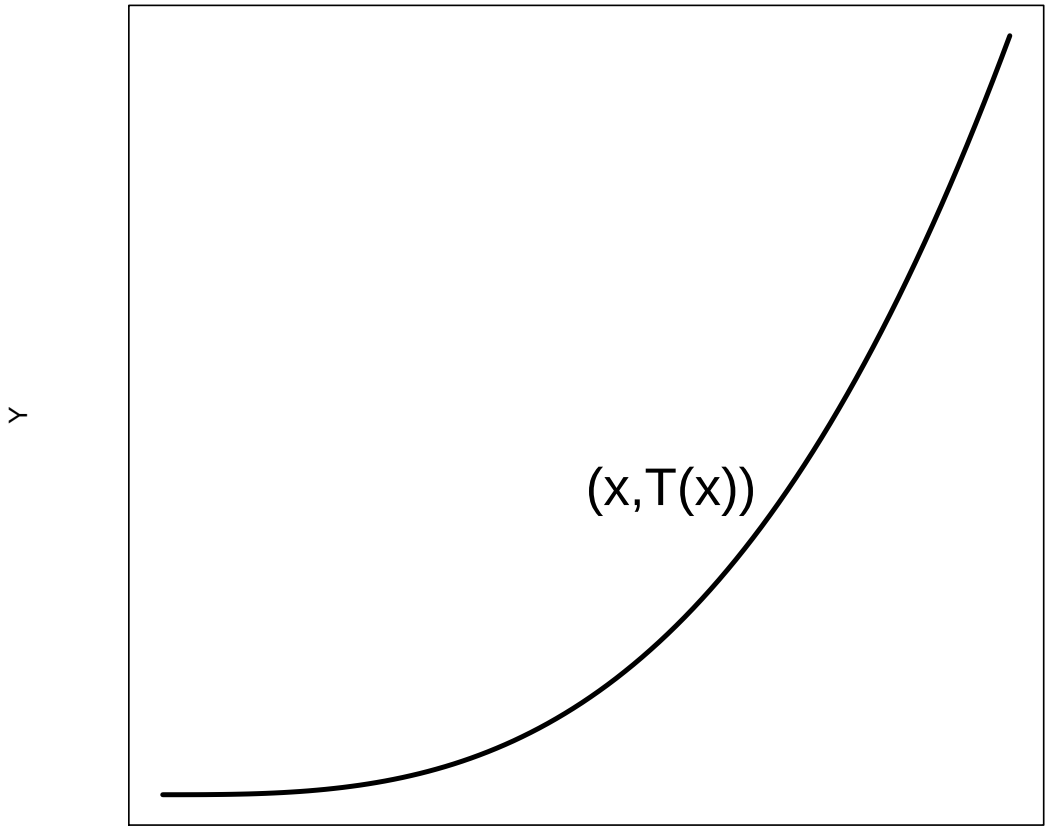
that is,

$P_s = F_{s\#}J$.

If a Transport Map Exists

If an optimal transport map T exists the the optimal coupling J is degenerate and is supported on the curve

$$\mathcal{S} = \{(x, T(x))\}$$



Regularized Optimal Transport

Find J (optimal transport plan) to minimize

$$\left(\int ||x - y||^p dJ(x, y) \right)^{1/p} + \lambda f(J)$$

for some f.

Regularized Optimal Transport

Find J (optimal transport plan) to minimize

$$\left(\int ||x - y||^p dJ(x, y) \right)^{1/p} + \lambda f(J)$$

for some f.

For example, Cuturi (2013) uses the entropy:

$$f(J) = - \int j(x, y) \log j(x, y)$$

Regularized Optimal Transport

Advantages:

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Disadvantages:

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Disadvantages:

(i) How to choose λ?

Regularized Optimal Transport

Advantages:

(i) fast algorithms (Sinkhorn-Knopp algorithm)

(ii) inference might be easier (Klatt, Tameling and Munk arXiv: 1810.09880)

Disadvantages:

(i) How to choose λ?

(ii) Effect of regularization is not clear.

Conclusion

Given two distributions P_0 and P_1 we can define an optimal transport map T.

Conclusion

Given two distributions P_0 and P_1 we can define an optimal transport map T.

We can define a geodesic (morphing) between two (or more) distributions.

Conclusion

Given two distributions P_0 and P_1 we can define an optimal transport map T.

We can define a geodesic (morphing) between two (or more) distributions.

Computation is expensive but doable.

Conclusion

Given two distributions P_0 and P_1 we can define an optimal transport map T.

We can define a geodesic (morphing) between two (or more) distributions.

Computation is expensive but doable.

Statistical inference (uncertainty about the estimated transport map) is in the works.