CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 22, 2021

Yale University

Last time

- Optimization on manifold
- Continuous time: Gradient flow
- Discrete time: Gradient descent and proximal gradient
- Strong convexity $\Rightarrow$ Gradient dominated
- Exponential convergence rate (with smoothness)

Today: Wasserstein metric and optimal transport

References

- Villani, *Topics in Optimal Transportation*, Springer, 2003
- Villani, *Optimal Transport: Old and New*, Springer, 2008
- Peyre & Cuturi, *Computational Optimal Transport*, Foundations and Trends in Machine Learning, 2019, available at `https://optimaltransport.github.io/book/`
- Peyre, *Course Notes on Computational Optimal Transport*, 2019, `https://optimaltransport.github.io/slides-peyre/CourseOT.pdf`

Space of distributions

Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote **space of probability distributions** over space $\mathcal{X}$

- $\mathcal{X} = \{1, \ldots, n\}$:

$$\mathcal{P}(\mathcal{X}) = \Delta_{n-1} = \{p \in \mathbb{R}^n \mid p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}$$

- $\mathcal{X}$ = graph, matroid, ...
- $\mathcal{X} = \mathbb{R}^n$, manifold, metric space, ...

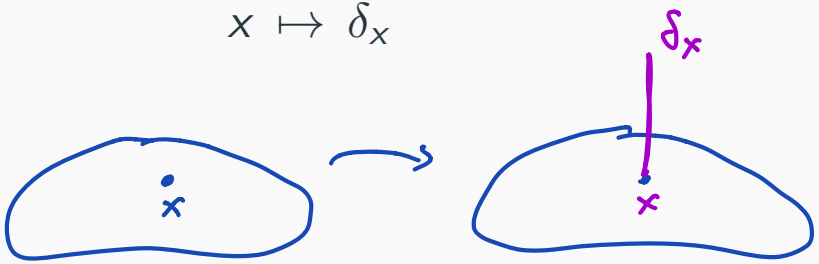
Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over space $\mathcal{X}$

- Modeling randomness on $\mathcal{X}$
- Includes deterministic:

$$\mathcal{X} \hookrightarrow \mathcal{P}(\mathcal{X})$$

$$x \mapsto \delta_x$$



Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over $\mathcal{X}$

- Dynamics on $\mathcal{X} \Rightarrow$ dynamics on $\mathcal{P}(\mathcal{X})$

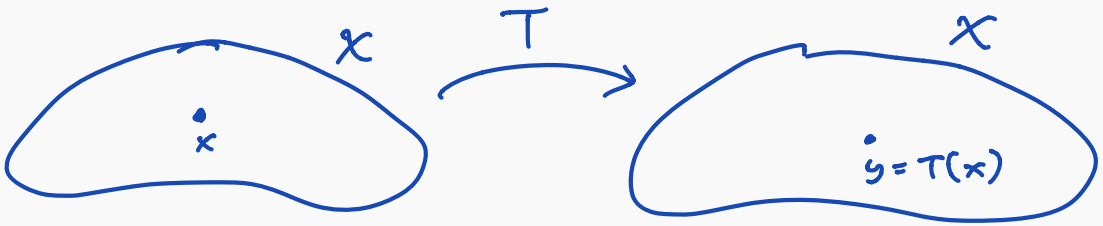
Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over $\mathcal{X}$

- Dynamics on $\mathcal{X} \Rightarrow$ dynamics on $\mathcal{P}(\mathcal{X})$
  - $\star$ Map $T : \mathcal{X} \to \mathcal{X}$ induces pushforward $T_{\#} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$

"If $X \sim \rho$, then $Y = T(X) \sim T_{\#}(\rho) = \nu$"

$\mathbb{P}_R(Y \in A) = \nu(A) = \rho(T^{-1}(A)) = \mathbb{P}_R(X \in T^{-1}(A))$



Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over $\mathcal{X}$

- Dynamics on $\mathcal{X} \Rightarrow$ dynamics on $\mathcal{P}(\mathcal{X})$
  - ⋆ Map $T : \mathcal{X} \to \mathcal{X}$ induces pushforward $T_{\#} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$

$$\text{If } X \sim \rho, \text{ then } Y = T(X) \sim T_{\#}(\rho) = \nu$$

$$\nu(A) = \rho(T^{-1}(A))$$

  - ⋆ Algorithm $x_{k+1} = T(x_k)$ on $\mathcal{X}$ induces algorithm $\rho_{k+1} = T_{\#}(\rho_k)$ on $\mathcal{P}(\mathcal{X})$

if $x_k \sim \rho_k$
and $x_{k+1} = T(x_k)$
then $x_{k+1} \sim \rho_{k+1} = T_{\#}(\rho_k)$

Space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over $\mathcal{X}$

- Dynamics on $\mathcal{X} \Rightarrow$ dynamics on $\mathcal{P}(\mathcal{X})$
  - ⋆ Map $T : \mathcal{X} \to \mathcal{X}$ induces pushforward $T_{\#} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$

    $$\text{If } X \sim \rho, \text{ then } Y = T(X) \sim T_{\#}(\rho) = \nu$$

    $$\nu(A) = \rho(T^{-1}(A))$$

  - ⋆ Algorithm $x_{k+1} = T(x_k)$ on $\mathcal{X}$ induces algorithm $\rho_{k+1} = T_{\#}(\rho_k)$ on $\mathcal{P}(\mathcal{X})$

- Stochastic dynamics on $\mathcal{X} \Rightarrow$ *deterministic* dynamics on $\mathcal{P}(\mathcal{X})$

  $T : \mathcal{X} \longrightarrow \mathcal{P}(\mathcal{X})$

Geometry on space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over space $\mathcal{X}$

- Many ways to measure distance/divergence:

$$\mathrm{TV}(\rho, \nu) = \frac{1}{2} \int_{\mathcal{X}} |\rho(x) - \nu(x)| \, dx$$

$$\chi^2_\nu(\rho) = \int_{\mathcal{X}} \nu(x) \left(\frac{\rho(x)}{\nu(x)} - 1 \right)^2 dx$$

$$H_\nu(\rho) = \int_{\mathcal{X}} \rho(x) \log \frac{\rho(x)}{\nu(x)} \, dx$$

work for any set $\mathcal{X}$

Geometry on space of distributions

Let $\mathcal{P}(\mathcal{X})$ denote space of probability distributions over space $\mathcal{X}$

- Many ways to measure distance/divergence:

$$\mathrm{TV}(\rho, \nu) = \frac{1}{2} \int_{\mathcal{X}} |\rho(x) - \nu(x)| \, dx$$

$$\chi^2_\nu(\rho) = \int_{\mathcal{X}} \nu(x) \left(\frac{\rho(x)}{\nu(x)} - 1 \right)^2 dx$$

$$H_\nu(\rho) = \int_{\mathcal{X}} \rho(x) \log \frac{\rho(x)}{\nu(x)} \, dx$$

- Want: geometry of $\mathcal{X} \Rightarrow$ geometry of $\mathcal{P}(\mathcal{X})$
- Assume $\mathcal{X} = \mathbb{R}^n$, also for $\mathcal{X} =$ manifold

Wasserstein metric

Wasserstein metric on the space of distributions $\mathcal{P}(\mathcal{X})$

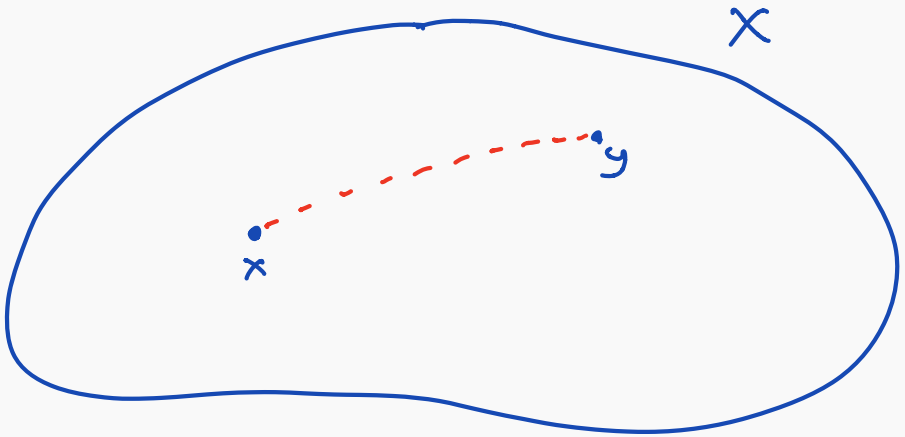
- Induced by *optimal transport* on $\mathcal{X}$
- Wasserstein W_2 metric: with quadratic distance cost function
- Natural metric, nice properties, connections to convex duality
- Smooth structure on $\mathcal{X} \Rightarrow$ smooth structure on $\mathcal{P}(\mathcal{X})$

Optimal transport

Cost function

Let $\mathcal{X}$ be space and $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *cost function*

$$c(x, y) = \text{ cost of transporting } x \text{ to } y$$

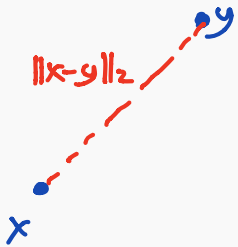


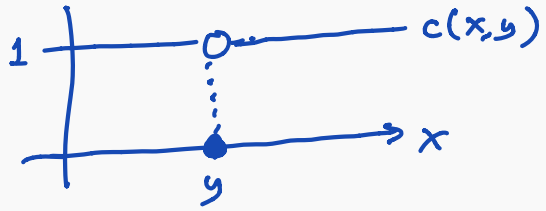
Cost function

Let $\mathcal{X}$ be space and $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *cost function*

$$c(x, y) = \text{ cost of transporting } x \text{ to } y$$

- $\mathcal{X} = \mathbb{R}^n$
  - $c(x, y) = \|x - y\|_2$
  - $c(x, y) = \|x - y\|_2^2$
  - $c(x, y) = \|x - y\|_2^p$
  - $c(x, y) = \phi(\|x - y\|_2)$ where $\phi\colon [0, \infty) \to \mathbb{R}$ convex, increasing
  - $c(x, y) = 1_{x \neq y} = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{else} \end{cases}$





Cost function

Let $\mathcal{X}$ be space and $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *cost function*

$$c(x, y) = \text{ cost of transporting } x \text{ to } y$$

- $\mathcal{X} = \mathbb{R}^n$
  - $c(x, y) = \|x - y\|_2$
  - $c(x, y) = \|x - y\|_2^2$
  - $c(x, y) = \|x - y\|_2^p$
  - $c(x, y) = \phi(\|x - y\|_2)$ where $\phi\colon [0, \infty) \to \mathbb{R}$ convex, increasing
  - $c(x, y) = 1_{x \neq y} = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{else} \end{cases}$
- $\mathcal{X} =$ manifold, metric space, ...
  - $c(x, y) = d(x, y)$, $d(x, y)^2$, ...

Transport cost

Let $\mathcal{X}$ be space and $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a cost function

The **transport cost** between $\rho, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\begin{aligned} \mathcal{T}_c(\rho, \nu) &= \inf_{\pi \in \Pi(\rho,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\, d\pi(x,y) \\ &= \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[c(X,Y)] \end{aligned}$$

$\mathcal{X}\times\mathcal{X}$

Ψ

- Infimum is over *couplings* π: joint distribution $(X, Y) \sim \pi$ with correct marginals $X \sim \rho$ and $Y \sim \nu$

Transport cost

Let $\mathcal{X}$ be space and $c\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a cost function

The **transport cost** between $\rho, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\begin{aligned} \mathcal{T}_c(\rho, \nu) &= \inf_{\pi \in \Pi(\rho,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)\, d\pi(x,y) \\ &= \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[c(X,Y)] \end{aligned}$$

- Infimum is over *couplings* π: joint distribution $(X, Y) \sim \pi$ with correct marginals $X \sim \rho$ and $Y \sim \nu$

- Optimization over conditional distributions (Markov chains) $P = (P_x \in \mathcal{P}(\mathcal{X})\colon x \in \mathcal{X})$ such that $\int_{\mathcal{X}} P_x(y)\, d\rho(x) = \nu(y)$

$\pi(x,y) = \rho(x) \cdot P(y \mid x)$

Earth mover's distance

$\mathcal{X} = \mathbb{R}^n$, Euclidean distance $c(x, y) = \|x - y\|_2$

Transport cost is **Earth-mover's distance**:

$$\mathcal{T}_c(\rho, \nu) = \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[\|X - Y\|_2] = W_1(\rho, \nu)$$

- Monge (1781): Transportation and allocation of resources
- Kantorovich (1938): Linear programming and minimax duality
  $\Rightarrow$ Nobel prize in Economics (1975)
- Computer vision (1989)

Total variation distance

Hamming distance:

$$c(x, y) = 1_{x \neq y} = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{else} \end{cases}$$

Transport cost is **Total Variation** distance:

$$\mathcal{T}_c(\rho, \nu) = \inf_{(X,Y)\sim\pi} \mathbb{P}(X \neq Y) = \mathrm{TV}(\rho, \nu)$$

$\mathbb{E}[1_{X \neq Y}] = \mathbb{E}[c(X, Y)]$

Wasserstein distance

Wasserstein distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^p$ for some $p \geq 1$

Let $\mathcal{P}_p(\mathbb{R}^n) = \{\rho \in \mathcal{P}(\mathbb{R}^n) \colon \mathbb{E}_\rho[\|X\|_2^p] < \infty\}$

Transport cost:

$$\mathcal{T}_p(\rho, \nu) = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|_2^p]$$

Wasserstein distance:

$$W_p(\rho, \nu) = \mathcal{T}_p(\rho, \nu)^{1/p}$$

Wasserstein distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^p$ for some $p \geq 1$

Let $\mathcal{P}_p(\mathbb{R}^n) = \{\rho \in \mathcal{P}(\mathbb{R}^n) \colon \mathbb{E}_\rho[\|X\|_2^p] < \infty\}$

Transport cost:

$$\mathcal{T}_p(\rho, \nu) = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|_2^p]$$

Wasserstein distance:

$$W_p(\rho, \nu) = \mathcal{T}_p(\rho, \nu)^{1/p}$$

Hölder's inequality: $p \leq q \quad \Rightarrow \quad W_p \leq W_q$

Wasserstein distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^p$ for some $p \geq 1$

Let $\mathcal{P}_p(\mathbb{R}^n) = \{\rho \in \mathcal{P}(\mathbb{R}^n) \colon \mathbb{E}_\rho[\|X\|_2^p] < \infty\}$

Transport cost:

$$\mathcal{T}_p(\rho, \nu) = \inf_{\pi \in \Pi(\rho, \nu)} \mathbb{E}_\pi[\|X - Y\|_2^p]$$

Wasserstein distance:

$$W_p(\rho, \nu) = \mathcal{T}_p(\rho, \nu)^{1/p}$$

Hölder's inequality: $p \leq q \quad \Rightarrow \quad W_p \leq W_q$

Theorem: W_p is a *distance metric* on $\mathcal{P}_p(\mathbb{R}^n)$

W_2 distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^2$ $(p = 2)$

W_2 **distance**:

$$W_2(\rho, \nu) = \sqrt{\mathcal{T}_2(\rho, \nu)} = \inf_{\pi \in \Pi(\rho, \nu)} \sqrt{\mathbb{E}_\pi[\|X - Y\|^2]}$$

W_2 distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^2$

W_2 distance:

$$W_2(\rho, \nu) = \sqrt{\mathcal{T}_2(\rho, \nu)} = \inf_{\pi \in \Pi(\rho, \nu)} \sqrt{\mathbb{E}_\pi[\|X - Y\|^2]}$$

- $\mathcal{X} \hookrightarrow \mathcal{P}(\mathcal{X})$ is an isometry:

$$W_2(\delta_x, \delta_y) = \|x - y\|_2$$

when $\rho = \delta_x$, $\nu = \delta_y$
$\Pi(\rho, \nu) = \{\delta_x \otimes \delta_y\}$

(whereas $\mathrm{TV}(\delta_x, \delta_y) = 1$ if $x \neq y$, and $\chi^2_\nu, H_\nu = \infty$)

$W_p(\delta_x, \delta_y) = \|x - y\|_2$

W_2 distance

$\mathcal{X} = \mathbb{R}^n$, $c(x, y) = \|x - y\|_2^2$

W_2 **distance**:

$$W_2(\rho, \nu) = \sqrt{\mathcal{T}_2(\rho, \nu)} = \inf_{\pi \in \Pi(\rho, \nu)} \sqrt{\mathbb{E}_\pi[\|X - Y\|^2]}$$

- $\mathcal{X} \hookrightarrow \mathcal{P}(\mathcal{X})$ is an isometry:

  $$W_2(\delta_x, \delta_y) = \|x - y\|_2$$

  (whereas $\mathrm{TV}(\delta_x, \delta_y) = 1$ if $x \neq y$, and $\chi^2_\nu, H_\nu = \infty$)

- $W_2(\rho, \delta_{x_0})^2 = \mathbb{E}_\rho[\|X - x_0\|^2]$

- If $\mu = \mathbb{E}_\rho[X]$, then

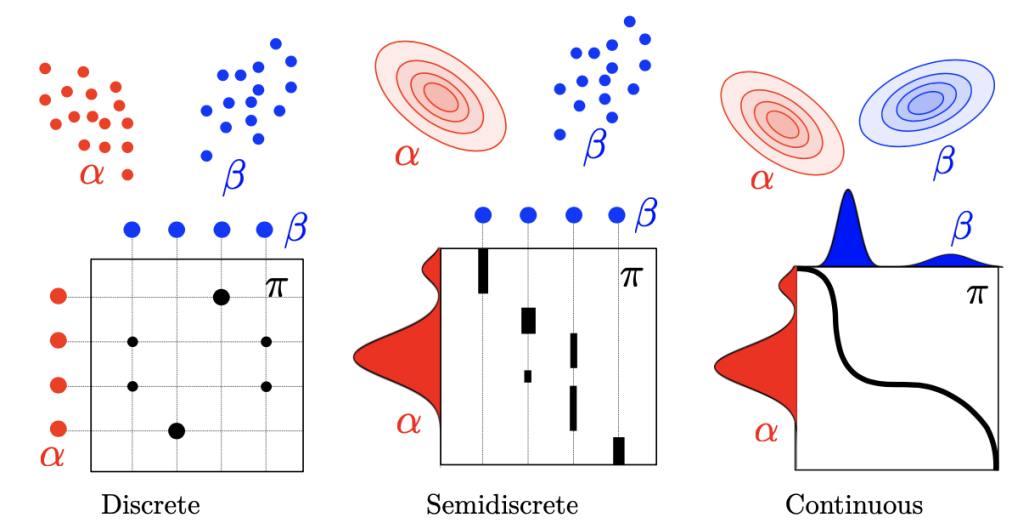
  $$W_2(\rho, \delta_\mu)^2 = \mathbb{E}_\rho[\|X - \mu\|^2] = \mathrm{Var}_\rho(X)$$

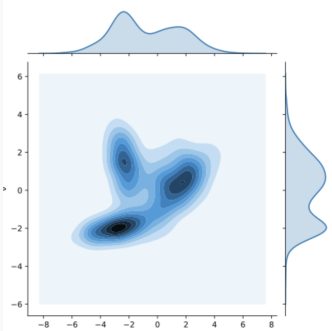
Optimal transport map

Optimal coupling

$$W_2(\rho,\nu) = \inf_{\pi\in\Pi(\rho,\nu)} \sqrt{\mathbb{E}_\pi[\|X-Y\|^2]}$$

- Existence and uniqueness of minimizer?





Optimal coupling

$$W_2(\rho,\nu)^2 = \inf_{\pi\in\Pi(\rho,\nu)} \mathbb{E}_\pi[\|X-Y\|^2]$$

Brenier's Theorem: Assume $\rho \ll dx$ on $\mathbb{R}^n$. Then there is a unique optimal coupling π, it is *deterministic*, and induced by the gradient of a *convex* function $\phi\colon \mathbb{R}^n \to \mathbb{R}$

$$d\pi(x,y) = d\rho(x)\,\delta_{\nabla\phi(x)}(y)$$

Optimal coupling

$$W_2(\rho,\nu)^2 = \inf_{\pi\in\Pi(\rho,\nu)} \mathbb{E}_\pi[\|X-Y\|^2]$$

Brenier's Theorem: Assume $\rho \ll dx$ on $\mathbb{R}^n$. Then there is a unique optimal coupling π, it is *deterministic*, and induced by the gradient of a *convex* function $\phi\colon \mathbb{R}^n \to \mathbb{R}$

$$d\pi(x,y) = d\rho(x)\,\delta_{\nabla\phi(x)}(y)$$

- Why convexity? Gradient of convex function is *monotone*
  Proof via Kantorovich duality, or cyclical monotonicity

$$W_2(\rho,\nu)^2 = \sup\left\{\mathbb{E}_\rho[\varphi(X)] + \mathbb{E}_\nu[\psi(Y)]\colon \varphi(x)+\psi(y) \le \|x-y\|^2\right\}$$

- For general cost c, use notions of c-convexity and c-duality

Optimal transport map

Let $\rho \ll dx$. For any ν, there is a unique convex $\phi\colon \mathbb{R}^n \to \mathbb{R}$ such that:

1. The gradient $\nabla\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ pushes ρ forward to ν:

$$X \sim \rho \quad \Rightarrow \quad Y = \nabla\phi(X) \sim \nu$$

2. $\nabla\phi$ is the **optimal transport map** (or *Brenier's map*) from ρ to ν:

$$W_2(\rho, \nu)^2 = \mathbb{E}_\rho[\|X - \nabla\phi(X)\|^2]$$

Optimal transport map

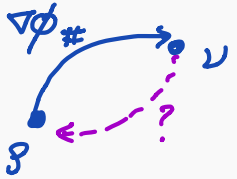
Let $\rho \ll dx$. For any ν, there is a unique convex $\phi\colon \mathbb{R}^n \to \mathbb{R}$ such that:

1. The gradient $\nabla\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ pushes ρ forward to ν:

$$X \sim \rho \quad \Rightarrow \quad Y = \nabla\phi(X) \sim \nu$$

2. $\nabla\phi$ is the **optimal transport map** (or *Brenier's map*) from ρ to ν:

$$W_2(\rho,\nu)^2 = \mathbb{E}_\rho[\|X - \nabla\phi(X)\|^2]$$



- If $\nu \ll dx$, the optimal transport map from ν to ρ is

$$\nabla\phi^* = (\nabla\phi)^{-1}$$

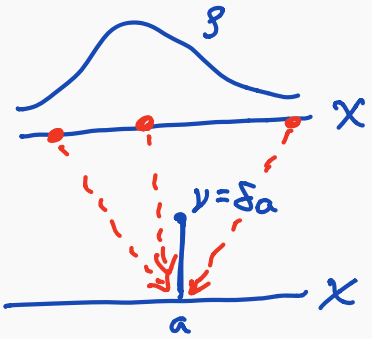
$y = \nabla\phi(x)$
$\Leftrightarrow\ x = \nabla\phi^*(y)$

where $\phi^*\colon \mathbb{R}^n \to \mathbb{R}$ is the *convex dual*: $\phi^*(y) = \sup_{x\in\mathbb{R}^n} \langle x, y\rangle - \phi(x)$

Examples

Example: Point mass

$$W_2(\rho,\nu) = \inf_{\pi\in\Pi(\rho,\nu)} \sqrt{\mathbb{E}_\pi[\|X-Y\|_2^2]}$$



- For $\rho \ll dx$, $\nu = \delta_a$. Optimal transport map:

$$\nabla\phi(x) = a$$

  from $\phi(x) = x^\top a$.

- Wasserstein distance: $W_2(\rho,\nu)^2 = \mathbb{E}_\rho[\|X-a\|_2^2]$

$$= \mathbb{E}_\rho\left[\|x - \nabla\phi(x)\|^2\right]$$

Example: Point mass

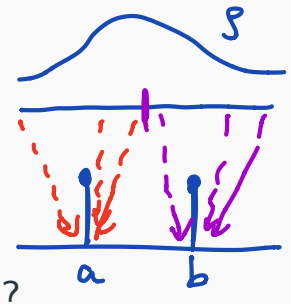
$$W_2(\rho, \nu) = \inf_{\pi \in \Pi(\rho,\nu)} \sqrt{\mathbb{E}_\pi[\|X - Y\|_2^2]}$$

- For $\rho \ll dx$, $\nu = \delta_a$. Optimal transport map:

  $$\nabla \phi(x) = a$$

  from $\phi(x) = x^\top a$.

- Wasserstein distance: $W_2(\rho, \nu)^2 = \mathbb{E}_\rho[\|X - a\|_2^2]$



- Exercise: What is optimal transport map to $\nu = \frac{1}{2}\delta_a + \frac{1}{2}\delta_b$?

Example: Gaussian

Let $\rho = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(\mu, \Sigma)$

- Recall if $X \sim \mathcal{N}(0, I)$, then $Y = \mu + \Sigma^{\frac{1}{2}} X \sim \mathcal{N}(\mu, \Sigma)$
- Optimal transport map:

$$\nabla \phi(x) = \mu + \Sigma^{\frac{1}{2}} x$$

from $\phi(x) = x^\top \mu + \frac{1}{2} x^\top \Sigma^{\frac{1}{2}} x$.

Example: Gaussian

Let $\rho = \mathcal{N}(0, I)$ and $\nu = \mathcal{N}(\mu, \Sigma)$

- Recall if $X \sim \mathcal{N}(0, I)$, then $Y = \mu + \Sigma^{\frac{1}{2}} X \sim \mathcal{N}(\mu, \Sigma)$
- Optimal transport map:

$$\nabla \phi(x) = \mu + \Sigma^{\frac{1}{2}} x$$

  from $\phi(x) = x^\top \mu + \frac{1}{2} x^\top \Sigma^{\frac{1}{2}} x$.

- Wasserstein distance:

$$W_2(\rho, \nu)^2 = \mathbb{E}[\|\mu + (\Sigma^{\frac{1}{2}} - I)X\|_2^2] = \|\mu\|_2^2 + \|\Sigma^{\frac{1}{2}} - I\|_{\text{HS}}^2$$

$$= \mathbb{E}\left[\|x - \nabla\phi(x)\|^2 \right]$$

Example: Gaussian

Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$

- Optimal transport map:

$$\nabla\phi(x) = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x-\mu_0)+\mu_1$$

- Wasserstein distance:

$$W_2(\rho,\nu)^2 = \|\mu_0-\mu_1\|_2^2 + \mathrm{Tr}(\Sigma_0+\Sigma_1-2(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}})$$

Example: Gaussian

Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$

- Optimal transport map:

$$\nabla\phi(x) = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x-\mu_0)+\mu_1$$

- Wasserstein distance:

$$W_2(\rho,\nu)^2 = \|\mu_0-\mu_1\|_2^2 + \mathrm{Tr}(\Sigma_0+\Sigma_1-2(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}})$$

- If $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$:

$$\nabla\phi(x) = \Sigma_0^{-\frac{1}{2}}\Sigma_1^{\frac{1}{2}}(x-\mu_0)+\mu_1$$

$$W_2(\rho,\nu)^2 = \|\mu_0-\mu_1\|_2^2 + \underbrace{\|\Sigma_0^{\frac{1}{2}}-\Sigma_1^{\frac{1}{2}}\|_{\mathrm{HS}}^2}_{\text{Buser metric}}$$

Example: 1-dimension

Let ρ, ν be probability distributions on $\mathbb{R}$ with cdf $F, G : \mathbb{R} \to [0,1]$

$F(x) = \mathbb{P}_\rho(X \le x)$

$G(y) = \mathbb{P}_\nu(Y \le y)$

- Optimal transport map from ρ to ν is

$$T = G^{-1} \circ F$$

  increasing $\Rightarrow$ derivative of convex function

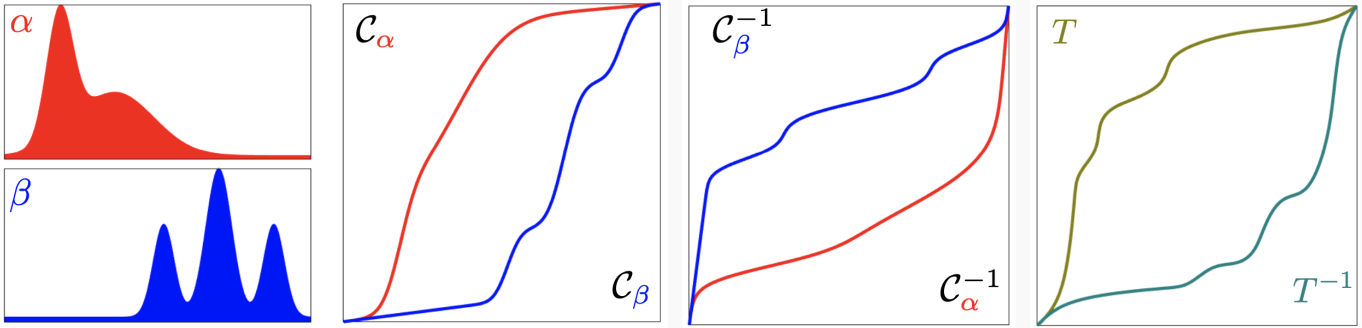
$X \sim \rho$ on $\mathbb{R}$

$\Rightarrow F(X) \sim$ Uniform on $[0,1]$

$\Rightarrow G^{-1}(F(x)) \sim \nu$ on $\mathbb{R}$

- Wasserstein distance is

$$W_2(\rho,\nu)^2 = \mathcal{T}_2(\rho,\nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 \, dt$$



Example: 1-dimension

Let ρ, ν be probability distributions on $\mathbb{R}$ with cdf $F, G : \mathbb{R} \to [0, 1]$

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

$$\mathcal{T}_2(\rho, \nu) = W_2(\rho, \nu)^2 = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 \, dt$$

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

$$\mathcal{T}_1(\rho, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)| \, dt = \int_{\mathbb{R}} |F(x) - G(x)| \, dx$$

- $c(x, y) = c(|x - y|)$, c convex and non-negative

$$\mathcal{T}_c(\rho, \nu) = \int_0^1 c(|F^{-1}(t) - G^{-1}(t)|) \, dt$$

- Kolmogorov-Smirnov:

$$\mathrm{KS}(\rho, \nu) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|$$

Bounds

Upper bound

Let ρ and ν with probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$

$$W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[\|X - Y\|^2]$$

- *Any* coupling $(X, Y) \sim \pi$ gives $W_2(\rho, \nu)^2 \leq \mathbb{E}_\pi[\|X - Y\|^2]$

Upper bound

Let ρ and ν with probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$

$$W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}_\pi[\|X - Y\|^2]$$

- *Any* coupling $(X, Y) \sim \pi$ gives $W_2(\rho, \nu)^2 \leq \mathbb{E}_\pi[\|X - Y\|^2]$
- Trivial coupling: $X \sim \rho$, $Y \sim \nu$ independent

$$\begin{aligned} W_2(\rho, \nu)^2 &\leq \|\mathbb{E}_\rho[X] - \mathbb{E}_\nu[Y]\|_2^2 + \mathrm{Var}_\rho(X) + \mathrm{Var}_\nu(Y) \\ &= \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2) \end{aligned}$$

Lower bound

Let ρ and ν with probability distributions on $\mathbb{R}^n$ with mean $\mu_1, \mu_2 \in \mathbb{R}^n$ and covariance $\Sigma_1, \Sigma_2 \succ 0$

Lower bound on Wasserstein distance:

$$W_2(\rho,\nu)^2 \geq \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1{}^{\frac{1}{2}}\Sigma_2\Sigma_1{}^{\frac{1}{2}})^{\frac{1}{2}})$$

If $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$, then

$$W_2(\rho,\nu)^2 \geq \|\mu_1 - \mu_2\|_2^2 + \|\Sigma_1{}^{\frac{1}{2}} - \Sigma_2{}^{\frac{1}{2}}\|_{\mathrm{HS}}^2$$

Geodesic

Displacement interpolation

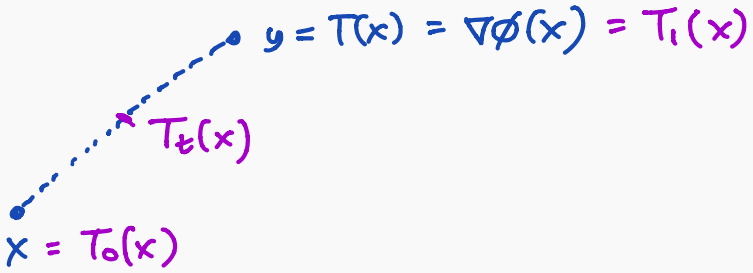
Let $\rho, \nu \ll dx$ be probability distributions on $\mathbb{R}^n$.

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν.

For $0 \leq t \leq 1$, define $T_t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$T_t(x) = (1-t)x + t\,T(x)$$

- Linear interpolation between $T_0(x) = x$ and $T_1(x) = T(x)$ in $\mathbb{R}^n$



Displacement interpolation

Let $\rho, \nu \ll dx$ be probability distributions on $\mathbb{R}^n$.

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν.

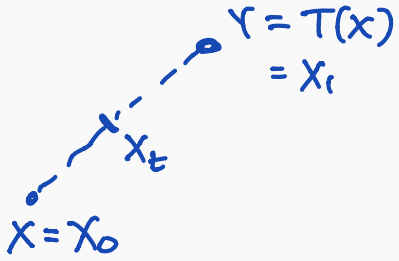
For $0 \le t \le 1$, define $T_t \colon \mathbb{R}^n \to \mathbb{R}^n$ by

$$T_t(x) = (1-t)x + t\,T(x)$$

- Linear interpolation between $T_0(x) = x$ and $T_1(x) = T(x)$ in $\mathbb{R}^n$

Define **displacement interpolation** $\rho_t \in \mathcal{P}(\mathbb{R}^n)$ by

$$\rho_t = (T_t)_{\#}\rho$$



- If $X \sim \rho$, then $X_t = (1-t)X + tT(X) \sim \rho_t$
- Interpolation between $\rho_0 = \rho$ and $\rho_1 = \nu$ in $\mathcal{P}(\mathbb{R}^n)$
- Also called *McCann's interpolation*

Displacement interpolation is geodesic

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν

Displacement interpolation: $\rho_t = (T_t)_{\#}\rho$

induced by linear interpolation $T_t(x) = (1-t)x + t\,T(x)$

Displacement interpolation is geodesic

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν

Displacement interpolation: $\rho_t = (T_t)_{\#}\rho$

induced by linear interpolation $T_t(x) = (1-t)x + t\,T(x)$

$= x + t\,(T(x) - x)$

Brenier's theorem: $T(x) = \nabla\phi(x)$ for some ϕ convex

$$\Rightarrow \quad T_t(x) = (1-t)x + t\,T(x)$$

$$= (1-t)x + t\,\nabla\phi(x)$$

$$= \nabla\left(\frac{(1-t)}{2}\|x\|^2 + t\,\phi(x)\right)$$

ϕ_t is also convex

$$T_t(x) = \nabla\phi_t(x)$$

$\Rightarrow$ Brenier's thm: T_t is the optimal transport map from ρ to ρ_t

$$\Rightarrow W_2(\rho, \rho_t)^2 = \mathbb{E}[\|x - T_t(x)\|^2]$$

$$= \mathbb{E}[\|x - ((1-t)x + t\,T(x))\|^2]$$

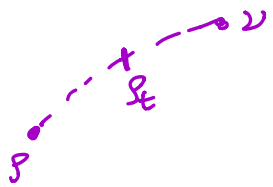
$$= \mathbb{E}[\|t\,(x - T(x))\|^2]$$

$$= t^2\,\mathbb{E}[\|x - T(x)\|^2]$$

$$= t^2\,\mathbb{E}[\|x - \nabla\phi(x)\|^2]$$

$$= t^2\,W_2(\rho, \nu)^2$$

$$\therefore \quad \boxed{W_2(\rho, \rho_t) = t\,W_2(\rho, \nu)}$$



Displacement interpolation is geodesic

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν

Displacement interpolation: $\rho_t = (T_t)_{\#}\rho$

induced by linear interpolation $T_t(x) = (1-t)x + t\,T(x)$

Brenier's theorem: $T(x) = \nabla\phi(x)$ for some ϕ convex

$\Rightarrow$ $T_t(x) = \nabla\phi_t(x)$ where $\phi_t(x) = \frac{(1-t)}{2}\|x\|^2 + t\phi(x)$ is convex

$\Rightarrow$ T_t is the optimal transport map from ρ to ρ_t

$\Rightarrow$ Wasserstein distance:

Displacement interpolation is geodesic

Let $T = \nabla\phi \colon \mathbb{R}^n \to \mathbb{R}^n$ be the optimal transport map from ρ to ν

Displacement interpolation: $\rho_t = (T_t)_{\#}\rho$

induced by linear interpolation $T_t(x) = (1-t)x + t\,T(x)$

Brenier's theorem: $T(x) = \nabla\phi(x)$ for some ϕ convex

$\Rightarrow$ $T_t(x) = \nabla\phi_t(x)$ where $\phi_t(x) = \frac{(1-t)}{2}\|x\|^2 + t\phi(x)$ is convex

$\Rightarrow$ T_t is the optimal transport map from ρ to ρ_t

$\Rightarrow$ Wasserstein distance:

$$W_2(\rho, \rho_t)^2 = \mathbb{E}_\rho[\|X - \nabla\phi_t(X)\|^2] = t^2 W_2(\rho, \nu)^2$$

$\Rightarrow$ Displacement interpolation is a **geodesic** in Wasserstein distance

Example: Gaussian

Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$

- Optimal transport map:

$$\nabla\phi(x) = A(x - \mu_0) + \mu_1$$

where $A = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}$

- Displacement interpolation: $\rho_t = \mathcal{N}(\mu_t, \Sigma_t)$ where

$$\mu_t = (1-t)\mu_0 + t\mu_1$$
$$\Sigma_t = ((1-t)I + tA)\Sigma_0((1-t)I + tA)$$

Example: Gaussian

Let $\rho = \mathcal{N}(\mu_0, \Sigma_0)$ and $\nu = \mathcal{N}(\mu_1, \Sigma_1)$

- Optimal transport map:

$$\nabla\phi(x) = A(x - \mu_0) + \mu_1$$

  where $A = \Sigma_0^{-\frac{1}{2}}(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}$

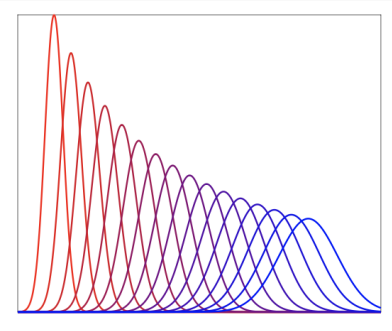
- Displacement interpolation: $\rho_t = \mathcal{N}(\mu_t, \Sigma_t)$ where

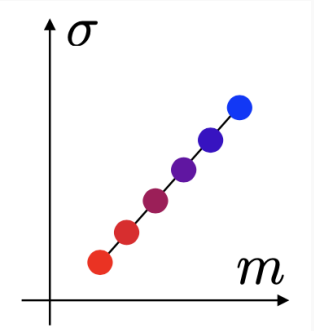
$$\mu_t = (1-t)\mu_0 + t\mu_1$$
$$\Sigma_t = ((1-t)I + tA)\Sigma_0((1-t)I + tA)$$

- If $\Sigma_0\Sigma_1 = \Sigma_1\Sigma_0$, then $\Sigma_t^{\frac{1}{2}} = (1-t)\Sigma_0^{\frac{1}{2}} + t\Sigma_1^{\frac{1}{2}}$

Example: Gaussian

Displacement interpolation between Gaussians stays Gaussian

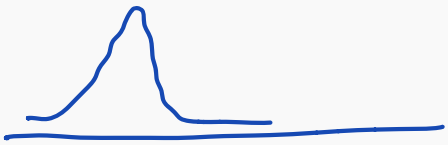


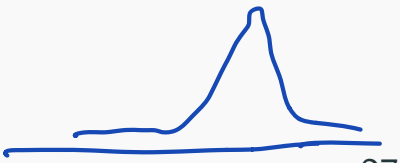


c.f. linear interpolation gives mixture of Gaussians

$$\tilde{\rho}_t = (1-t)\rho + t\nu$$

$$= (1-t)\mathcal{N}(\mu_1, \Sigma_1) + t\mathcal{N}(\mu_2, \Sigma_2)$$



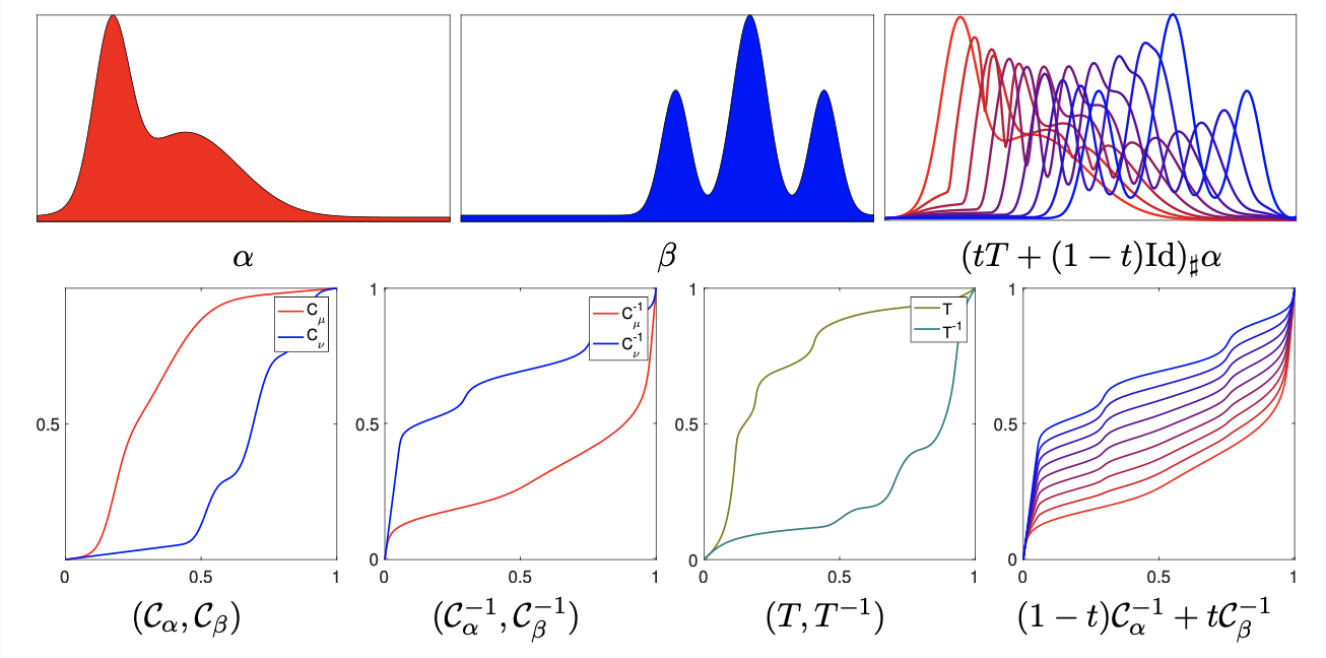


Example: 1-dimension

Let ρ, ν be probability distributions on $\mathbb{R}$ with cdf $F_0, F_1 : \mathbb{R} \to [0, 1]$

Displacement interpolation ρ_t has cdf F_t where

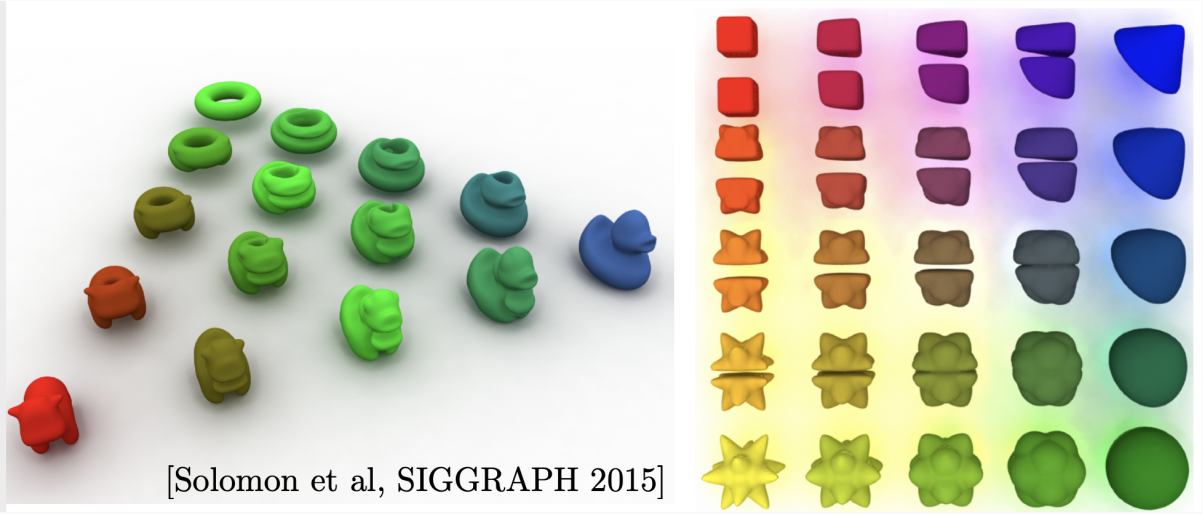
$$F_t^{-1} = (1-t)F_0^{-1} + tF_1^{-1}$$



Wasserstein barycenter

Average of distributions $(\rho_i)_{i=1}^m$ with weights $\sum_{i=1}^m \lambda_i = 1$

$$\rho^* = \arg \min_{\rho \in \mathcal{P}(\mathcal{X})} \sum_{i=1}^m \lambda_i W_2(\rho, \rho_i)^2$$



[Solomon et al, SIGGRAPH 2015]