UC San Diego

Embedding Learning by Optimal Transport

Presenter: An Yan, Fanbo Xiang, Yiming Zhang

March 05, 2019

Outline

- Wasserstein Distance
  - Optimal Transport
  - Exact Algorithm
- Learning Wasserstein Embeddings
- Entropic Transport
  - Entropic Regularization
  - Sinkhorn Divergence
- Learning Entropic Wasserstein Embeddings

Review: Optimal Transport

Discrete Kantorovish formulation(Earth mover's distance)

Discrete distributions $\mathbf{a} \in \mathbb{R}_+^n$, $\mathbf{b} \in \mathbb{R}_+^m$. Cost matrix $\mathbf{C} \in \mathbb{R}_+^{n \times m}$. $\mathbf{C}_{i,j}$ denotes the unit cost of transporting mass from ith point in $\mathbf{a}$ to jth point in $\mathbf{b}$.

$$\mathbf{U}(a, b) = \{\mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P}\mathbb{1}_m = \mathbf{a}, \mathbf{P}^T \mathbb{1}_n = \mathbf{b}\}$$

$\mathbf{P}_{i,j}$ denotes how much mass from ith point in $\mathbf{a}$ is transported to the jth point in $\mathbf{b}$. $\mathbf{U}(a, b)$ is all valid transport plans. $\mathbf{P}$ is known as a coupling matrix.

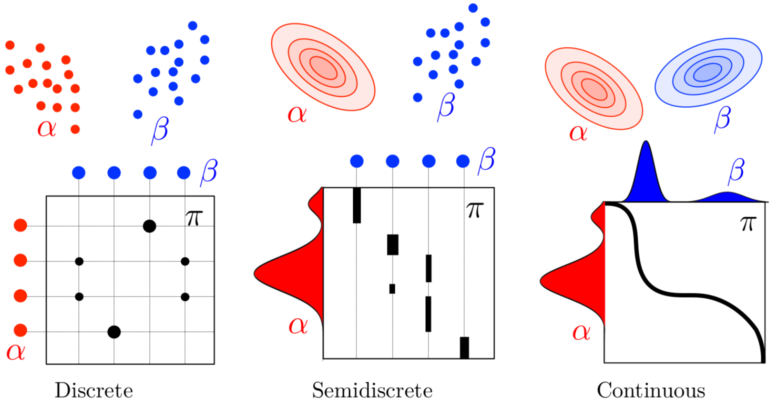
(Discrete) Optimal transport

A transport plan is optimal if it has the lowest cost.

$$L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}$$

Review: Optimal Transport

Moving mass from 1 distribution to the other.



Review: Optimal Transport

General formulation

$$\mathcal{L}_C(\alpha, \beta) = \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y)$$

Probabilistic interpretation

$$\mathcal{L}_C(\alpha, \beta) = \min_{X,Y} \{\mathbb{E}(c(X, Y)) : X \sim \alpha, Y \sim \beta\}$$

Intuition

Optimal transport gives a distance measure between probability distributions.

Wasserstein Distance

A special case of optimal transport. “A natural way to lift ground distance to distribution distance.”

Definition

Let $P_p(\Omega)$ be the set of Borel probability measures with finite pth moment defined on a given metric space (Ω, d). The p-Wasserstein metric W_p, for $p \geq 1$, on $P_p(\Omega)$ between distribution μ and ν, is defined as

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

1-Wasserstein Distance

Primal Problem

$$KP(\mu,\nu) = \min_{\gamma} \int_{\Omega\times\Omega} d(x,y)d\gamma(x,y)$$

$$s.t. \quad \int_Y d\gamma(x,y) = p(x), \int_X d\gamma(x,y) = q(y)$$

$$\gamma(x,y) \geq 0$$

Kantorovich-Rubinstein theorem

$$DP(\mu,\nu) = \max_{\phi\in Lip_1(X)} \int_X \phi(x)p(x)dx - \int_X \phi(x)q(x)dx$$

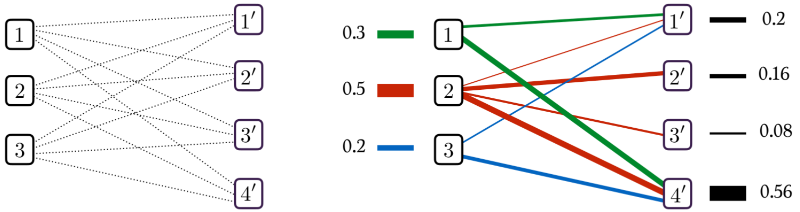
$$DP(\mu,\nu) = \max_{\phi\in Lip_1(X)} \mathbb{E}_p\phi(x) - \mathbb{E}_q\phi(x)$$

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

Algorithm for Optimal Transport

Discrete problem: linear programming

Can be formulated as a minimum cost maximum flow problem.



Any Questions?

Learning Wasserstein Embeddings

Idea

- Treat each data point as a distribution.
- Consider p-Wasserstein distance between data points.
- Embed Wasserstein space into Euclidean space.
- Learn this embedding with a neural network.

Entropic Regularization

Kantorovish formulation

$$U(a,b) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} : \mathbf{P}\mathbb{1}_m = \mathbf{a}, \mathbf{P}^T\mathbb{1}_n = \mathbf{b}\}$$

$\mathbf{P}_{i,j}$ denotes how much mass from ith point in $\mathbf{a}$ is transported to the jth point in $\mathbf{b}$. $U(a,b)$ is all valid transport plans. $\mathbf{P}$ is known as a coupling matrix.

Entropy

Discrete entropy of a coupling matrix $\mathbf{P}$:

$$\mathbf{H}(\mathbf{P}) := -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$$

$\mathbf{H}(\mathbf{P}) = -\infty$ if any entry of $\mathbf{P}$ is negative or 0.

Entropic Regularization

property

$\mathbf{H}$ is 1-strongly concave:

$$\forall x, y, (\nabla f(x) - \nabla f(y))^T (x - y) \leq ||x - y||_2^2$$

$$\forall x, -Hf(x) - I \text{ is positive semidefinite}$$

idea

Larger $\mathbf{H}(\mathbf{P}) \rightarrow$ distribution more uniform.
We can use $\mathbf{H}$ to regularize optimal transport.

$$L_{\mathbf{c}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle$$

$$L_{\mathbf{c}}^{\epsilon}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \epsilon \mathbf{H}(\mathbf{P})$$

Entropic Regularization

$$L_{\mathbf{c}}^{\epsilon}(\mathbf{a},\mathbf{b}) = \min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \epsilon \mathbf{H}(\mathbf{P})$$

$L_{\mathbf{c}}^{\epsilon}(\mathbf{a},\mathbf{b})$ is known as the **Sinkhorn divergence**.

Properties

1. There exists unique solution $\mathbf{P}_\epsilon$.
2. When $\epsilon \to 0$, $\mathbf{P}_\epsilon \to \mathbf{P}$.
3. When $\epsilon \to \infty$, $\mathbf{P}_\epsilon \to \mathbf{a}\mathbf{b}^T$ (uniform distribution).

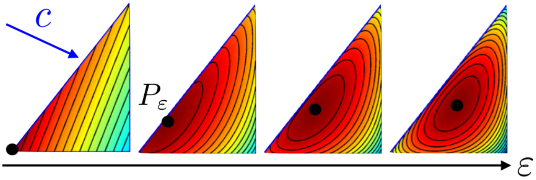


Figure 4.1: Impact of ε on the optimization of a linear function on the simplex, solving $\mathbf{P}_\varepsilon = \operatorname{argmin}_{\mathbf{P}\in\Sigma_3} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon \mathbf{H}(\mathbf{P})$ for a varying ε.

Entropic Regularization

Proposition (4.3)

Solution to the discrete entropic optimal transport problem

$$L_{\mathbf{C}}^{\epsilon}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \epsilon \mathbf{H}(\mathbf{P})$$

is unique and has the form

$$\forall (i, j) \in [n] \times [m], \mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j$$

or equivalently,

$$\mathbf{P} = \mathrm{diag}(\mathbf{u}) \mathbf{K} \mathrm{diag}(\mathbf{v})$$

where

$$\mathbf{K}_{i,j} = e^{-\mathbf{C}_{i,j}/\epsilon}, (\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$$

Entropic Regularization

Sinkhorn iterations

$$\mathbf{P} = \text{diag}(\mathbf{u})\mathbf{K}\text{diag}(\mathbf{v})$$

Adding constraints $\mathbf{P}\mathbb{1}_m = \mathbf{a}, \mathbf{P}^T\mathbb{1}_n = \mathbf{b}$,

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a}, \mathbf{v} \odot (\mathbf{K}^T\mathbf{u}) = \mathbf{b}$$

This problem is known as "matrix scaling" and can be solved iteratively:

$$\mathbf{u}^{(l+1)} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(l)}}, \mathbf{v}^{(l+1)} = \frac{\mathbf{b}}{\mathbf{K}^T\mathbf{u}^{(l+1)}}$$

Note: this algorithm converges but possibly to different values for different initialization, since $(\lambda\mathbf{u}, \mathbf{v}/\lambda)$ is also a solution.

Entropic Regularization

Complexity

Let $n = m$ for simplicity, to achieve approximate transport plan $\hat{\mathbf{P}} \in U(\mathbf{a}, \mathbf{b})$ with $\langle \hat{\mathbf{P}}, \mathbf{C} \rangle \leq L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) + \tau$, the time complexity is

$$O(n^2 \log n \tau^{-3})$$

Remarks

The Sinkhorn iteration approximates optimal transport. Given enough time, it can give arbitrarily close approximations.

Any Questions?

Learning Entropic Wasserstein Embeddings

Idea

- Want "similar" data points to be close in a embedding space.
- Use a Wasserstein space as the embedding space.
- Use Sinkhorn iteration as a layer in the neural network.