

Wasserstein-Distance

This is an brief introduction to Wasserstein-Distance, including its formulation, computation and application.

Tutorials

1. [Optimal Transport for Applied Mathematicians](#)
2. [Optimal Transport for Data Analysis](#)
3. [A user's guide to optimal transport](#)

Introduction

We will start from some some intuitive examples.

Existing metrics for evaluating distance between distributions

1. metrics

1. KL divergence: $D_{\text{KL}}(P\|Q) = -\sum_i P(i) \ln \frac{Q(i)}{P(i)}$
2. JS divergence: $D_{\text{JS}}(P, Q) = \frac{1}{2} \left(D_{\text{KL}} \left(P\| \frac{P+Q}{2} \right) + D_{\text{KL}} \left(Q\| \frac{P+Q}{2} \right) \right)$
3. F divergence: $D_f(p\|q) = \int q(x) f \left(\frac{p(x)}{q(x)} \right) dx$, where f is a convex function.

2. Issues

1. can not evaluate 2 distributions with different support set.
 1. example: $\{p(x)|x \in [0, 1]\}$ and $\{q(y)|y \in [2, 3]\}$
2. use KL/JS divergence as loss function -> gradient vanishing!
3. need well-defined distance metric

Transportation problem

	D_1	D_2	\dots	D_n	Supply
O_1	c_{11}	c_{12}	\dots	c_{1n}	a_1
O_2	c_{21}	c_{22}	\dots	c_{2n}	a_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
O_m	c_{m1}	c_{m2}	\dots	c_{mn}	a_m
Demand	b_1	b_2	\dots	b_n	

1. Problem:

1. origin: $\{O_1, \dots, O_m\}$, destination: $\{D_1, \dots, D_n\}$
2. supply: $\{a_1, \dots, a_m\}$, demand: $\{b_1, \dots, b_n\}$
3. transport supply goods in origin locations to satisfy demand in destinations
4. transport plan/matrix: x_{ij}
5. transport cost: c_{ij}

2. Formulation

1. Primal formation

$$\text{Minimize} \quad \sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij}$$

Subject to:

$$\begin{aligned} \sum_{j=1}^n x_{ij} &= a_i & \text{for } i = 1, 2, \dots, m \\ \sum_{i=1}^m x_{ij} &= b_j & \text{for } j = 1, 2, \dots, n \\ x_{ij} &\geq 0 & \text{for } i = 1, 2, \dots, m \text{ and } j = 1, 2, \dots, n \end{aligned}$$

2. Dual formulation

$$\text{Maximize} \quad \sum_{i=1}^m a_i f_i + \sum_{j=1}^n b_j g_j$$

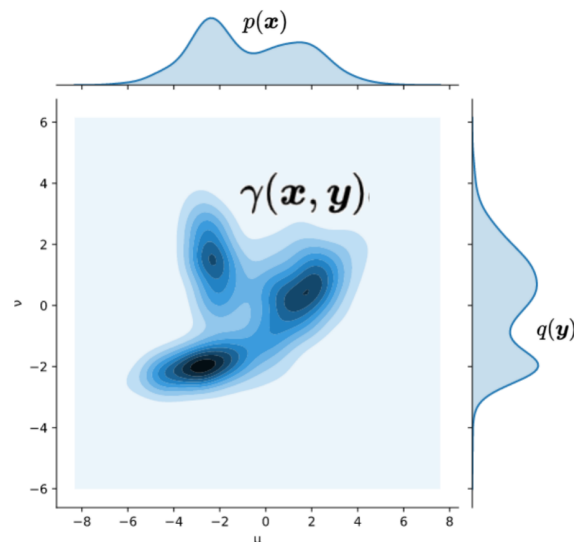
Subject to:

$$f_i + g_j \leq c_{ij} \quad \text{for } i = 1, 2, \dots, m, \text{ for } j = 1, 2, \dots, n$$

Formulation

1. Wasserstein distance (Kantorovich formulation)

1. view it as a continuous version of transportation problem



$$\text{Minimize } \mathcal{W}[p, q] = \inf_{\gamma \in \Pi[p, q]} \iint \gamma(x, y) c(x, y) dx dy$$

Subject to:

$$\int \gamma(x, y) dy = p(x)$$

$$\int \gamma(x, y) dx = q(y)$$

2. cost function $c(x, y)$:

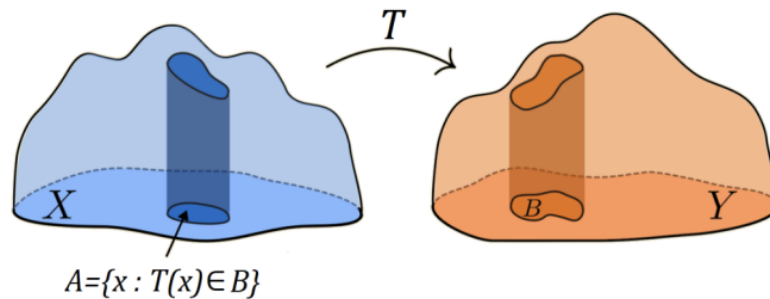
1. any norm, $\|x - y\|_1$, $\|x - y\|_2$, $\|x - y\|_2^2$

3. joint distribution $\gamma(x, y)$:

1. with marginal distribution $\gamma(x) = p(x)$, $\gamma(y) = q(y)$

2. Optimal transport and Wasserstein distance

1. Optimal transport (Monge formulation)



$$C_M(T) = \int_{\Omega} c(x, T(x)) p(x) dx$$

$$q = T(p)$$

1. transport map: $q(y) = T(p(x))$

1. non-linear constraint

3. Several dual formulations

1. Kantorovich-Rubinstein Duality

$$\mathcal{W}[p, q] = \max_{f, g} \left\{ \int [p(x)f(x) + q(x)g(x)] dx \mid f(x) + g(y) \leq c(x, y) \right\}$$

1. primal-dual optimality condition

1. $f(x) + g(y) = c(x, y)$

2. Proof:

1. forward: if primal and dual reach optimality, then

$$\underbrace{\iint \gamma(x, y) c(x, y) dx dy}_{\text{primal formulation}} \quad (1)$$

$$= \underbrace{\int [p(x) f(x) + q(x) g(x)] dx}_{\text{primal=dual when reaching optimality}} \quad (2)$$

$$= \underbrace{\iint [f(x) + g(y)] \gamma(x, y) dx dy}_{\text{marginal distribution}} \quad (3)$$

$$\rightarrow f(x) + g(y) = c(x, y) \quad (4)$$

2. backward: if $f(x) + g(y) = c(x, y)$ holds, then:

$$\underbrace{\int [p(x) f(x) + q(x) g(x)] dx}_{\text{dual formulation}} \quad (5)$$

$$= \underbrace{\iint [f(x) + g(y)] \gamma(x, y) dx dy}_{\text{marginal distribution}} \quad (6)$$

$$= \underbrace{\iint \gamma(x, y) c(x, y) dx dy}_{f(x)+g(y)=c(x,y)} \quad (7)$$

$$\rightarrow \text{primal} = \text{dual when reaching optimality} \quad (8)$$

2. Lipschitz constrained formulation

$$\mathcal{W}[p, q] = \max_f \left\{ \int [p(x) f(x) - q(x) f(x)] dx \mid \|f\|_L \leq 1 \right\}$$

1. consider the optimality condition when $x = y$

$$1. f(y) + g(y) = c(y, y) = 0 \rightarrow g(y) = -f(y)$$

2. take $g(y) = -f(y)$ into the $\mathcal{W}[p, q]$

$$1. \text{ objective function: } \max_f \left\{ \int [p(x) f(x) - q(x) f(x)] dx \right\}$$

$$2. \text{ constraints: } \|f\|_L \leq 1$$

$$1. f(x) - f(y) \leq c(x, y) \text{ and } f(y) - f(x) \leq c(y, x)$$

$$2. \|f\|_L = \frac{|f(x) - f(y)|}{c(x, y)} \leq 1$$

3. Unconstrained formulation

$$\mathcal{W}[p, q] = \max_f \int f(x) dp(x) + \int \min_x [c(x, y) - f(x)] dq(y)$$

1. C-transform:

For $f \in C(\Omega)$ define its c -transform $f^c \in C(\Omega)$ by

$$f^c(y) = \inf\{c(x, y) - f(x) \mid x \in \Omega\}$$

and its \bar{c} -transform $g^{\bar{c}} \in C(\Omega)$ by

$$g^{\bar{c}}(x) = \inf\{c(x, y) - g(y) \mid y \in \Omega\}$$

1. $f^{\hat{c}}(x) \geq f(x)$, "=" holds when f is concave
2. Consider $g(y)$ is the C-transform of $f(x)$
 1. $f^c(y) = \inf_x \{c(x, y) - f(x)\}$
 2. Proof of such a transform will not affect the optimality
 1. prove $f(x)$ and $f^c(y)$ satisfy the constraint

$$f(x) + \inf\{c(x, y) - f(x)\} \tag{9}$$

$$\leq f(x) + c(x, y) - f(x) \tag{10}$$

$$= c(x, y) \tag{11}$$

The constraint is always be satisfied under C-transform

2. prove $f(x)$ and $f^c(y)$ reach optimality condition

$$f(x) = g^c(x) \tag{12}$$

$$\rightarrow f^c(y) = g^{\hat{c}}(y) \geq g(y) \tag{13}$$

$$\rightarrow f(x) + f^c(y) \geq f(x) + g(y) \tag{14}$$

when $f(x) + g(y) = c(x, y)$, $c(x, y) \leq f(x) + f^c(y) \leq c(x, y)$

Therefore $f(x) + f^c(y) = c(x, y)$ and reaches optimality.

4. Quadratic cost function

1. quadratic cost function: $c(x, y) = \frac{1}{2}\|x - y\|^2$
2. The C-transform can be simplified as:

$$\begin{aligned}
f(x) &= \inf_y \left\{ \frac{1}{2} \|x - y\|^2 - g(y) \right\} \\
&= \frac{1}{2} \|x\|^2 + \inf_y \left\{ -\langle x, y \rangle + \frac{1}{2} \|y\|^2 - g(y) \right\} \\
&= \frac{1}{2} \|x\|^2 - \underbrace{\sup_y \left\{ \langle x, y \rangle - \left[\frac{1}{2} \|y\|^2 - g(y) \right] \right\}}_{:=\phi(x): \text{convex}}
\end{aligned}$$

1. $\phi(x)$ is the convex conjugate of $\frac{1}{2} \|y\|^2 - g(y)$
3. Brenier theorem:
 1. Under quadratic case, optimal transport map $T(x)$ is equivalent with transport plan $\gamma(x, y)$

$$T(x) = x - \nabla f(x) = x - (x - \nabla \phi(x)) = \nabla \phi(x)$$

5. Convex formulation

1. under quadratic case:

$$\begin{aligned}
f(x) + g(y) &\leq \frac{1}{2} \|x - y\|_2^2 \iff \\
\left[\frac{1}{2} \|x\|_2^2 - f(x) \right] + \left[\frac{1}{2} \|y\|_2^2 - g(y) \right] &\geq \langle x, y \rangle
\end{aligned}$$

2. define:

1. $f'(x) = \frac{1}{2} \|x\|_2^2 - f(x)$
2. $g'(y) = \frac{1}{2} \|y\|_2^2 - g(y)$

3. The objective function becomes:

$$\begin{aligned}
\mathcal{W}[p, q] &= C_{p,q} - \min_{f', g'} \{ \mathbb{E}_p[f'(x)] + \mathbb{E}_q[g'(y)] \mid f'(x) + g'(y) \geq \langle x, y \rangle \} \\
C_{p,q} &= \frac{1}{2} \mathbb{E}_p[\|X\|_2^2] + \mathbb{E}_q[\|Y\|_2^2]
\end{aligned}$$

4. apply the conjugate transformation

$$1. \ g'(y) = f'^*(y) = \sup_x \left\{ \langle x, y \rangle - \underbrace{\left[\frac{1}{2} \|x\|^2 - f(x) \right]}_{f'(x)} \right\}$$

5. unconstrained optimization

$$\mathcal{W}[p, q] = C_{p,q} - \min_{f', g'} \left\{ \mathbb{E}_p[f'(x)] + \mathbb{E}_q[f'^*(y)] \right\}$$

1. similar proof as C-transform

1. constraint:

$$1. \ f'(x) + f'^*(y) \geq \langle x, y \rangle$$

2. optimality

1. $f^{**} \leq f$, "=" holds when f is convex
2. f' and g' are convex

6. According to the Brenier theorem, when reach optimality

1. $\nabla g'(y) = T(y)$ is the optimal transport map
2. $f'^*(y) = \sup_x \left\{ \langle x, y \rangle - \left[\frac{1}{2} \|x\|^2 - f(x) \right] \right\} = \langle T(y), y \rangle - \left[\frac{1}{2} \|T(y)\|^2 - f(T(y)) \right]$

7. convex formulation:

$$\mathcal{W}[p, q] = C_{p,q} - \min_{f' \in \text{cvx}} \max_{g' \in \text{cvx}} \left\{ \mathbb{E}_p[f'(x)] + \mathbb{E}_q[f'^*(y)] \right\}$$

Computation

Application

Computation

Discretization

Minimize $\sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij}$

Subject to:

$$\begin{aligned} \sum_{j=1}^n x_{ij} &= a_i & \text{for } i = 1, 2, \dots, m \\ \sum_{i=1}^m x_{ij} &= b_j & \text{for } j = 1, 2, \dots, n \\ x_{ij} &\geq 0 & \text{for } i = 1, 2, \dots, m \text{ and } j = 1, 2, \dots, n \end{aligned}$$

1. Linear programming

2. Sinkhorn iteration

1. [Sinkhorn Distances: Lightspeed Computation of Optimal Transport](#)

3. ADMM

1. [Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support](#)
2. [A Fast Globally Linearly Convergent Algorithm for the Computation of Wasserstein Barycenters](#)

Lipschitz constraint

$$\mathcal{W}[p, q] = \max_f \left\{ \int [p(x)f(x) - q(x)f(x)] dx \mid \|f\|_L \leq 1 \right\}$$

1. Wasserstein Generative Adversarial Networks

2. Improved Training of Wasserstein GANs

C-transform

1. Three-Player Wasserstein GAN via Amortised Duality

Quadratic + ICNN

1. 2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs
2. Optimal transport mapping via input convex neural networks
3. Input Convex Neural Networks
4. Optimal Control Via Neural Networks: A Convex Approach

2-step Computation

1. A Two-Step Computation of the Exact GAN Wasserstein Distance
2. Wasserstein GAN with Quadratic Transport Cost

GMM

1. [Optimal transport for Gaussian mixture models](#)

Empirical study

1. Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark