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

1. metrics
    1. KL divergence: $D_{\mathrm{KL}}(P\|Q) = -\sum_i P(i) \ln \frac{Q(i)}{P(i)}$
    2. JS divergence: $D_{\mathrm{JS}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(P\|\frac{P+Q}{2}\right) + D_{\mathrm{KL}}\left(Q\|\frac{P+Q}{2}\right)\right)$
    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	$\cdots$	D_n	Supply
O_1	c_{11}	c_{12}	$\cdots$	c_{1n}	a_1
O_2	c_{21}	c_{22}	$\cdots$	c_{2n}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
O_m	c_{m1}	c_{m2}	$\cdots$	c_{mn}	a_m
Demand	b_1	b_2	$\cdots$	b_n	

1. Problem:
    1. origin: $\{O_1, ..., O_m\}$, destination: $\{D_1, ..., D_n\}$
    2. supply: $\{a_1, ..., a_m\}$, demand: $\{b_1, ..., b_n\}$
    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
       Minimize $\sum_{i=1}^m \sum_{j=1}^n c_{ij} x_{ij}$
       Subject to:

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

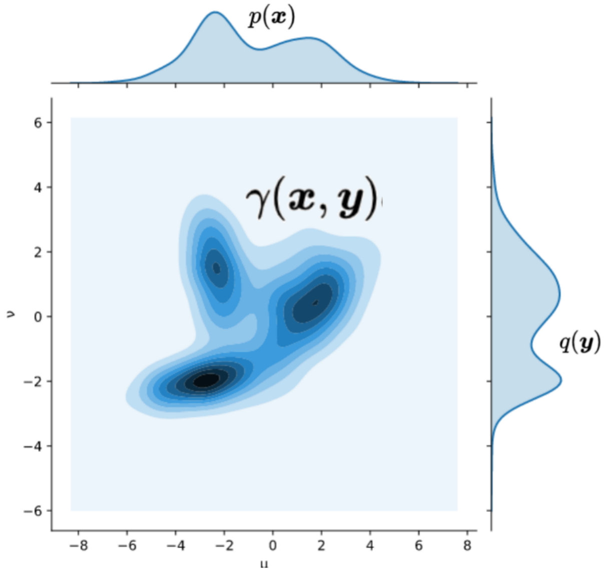
    2. Dual formulation
       Maximize $\sum_{i=1}^m a_i f_i + \sum_{i=1}^n b_i g_i$
       Subject to:

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

Formulation

Wasserstein distance (Kantorovich formulation)

1. view it as a continuous version of transportation problem



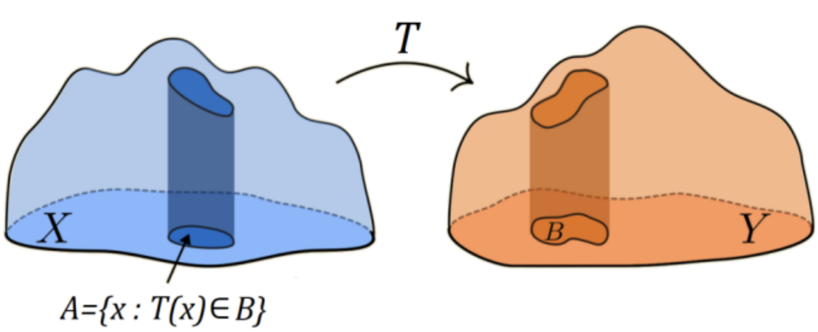
Minimize $\mathcal{W}[p,q] = \inf_{\gamma \in \Pi[p,q]} \iint \gamma(x,y)c(x,y)dxdy$
Subject to:

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

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 $\gamma(x) = p(x), \gamma(y) = q(y)$

Optimal transport and Wasserstein distance

1. Optimal transport (Monge formulation)



$$C_M(T) = \int_\Omega c(x, T(x))p(x)\mathrm{d}$$
$$q = T(p)$$

1. transport map: $q(y) = T(p(x))$
    1. non-linear constraint

Several dual formulations

Kantorovich-Rubinstein Duality

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

1. primal-dual optimality condition
1. $f(x)+g(y)=c(x,y)$
2. Proof:
1. forward: if primal and dual reach optimality, then

$$\begin{aligned}
&\underbrace{\iint \gamma(x,y)c(x,y)dxdy}_{\text{primal formulation}} && (1)\\
&= \underbrace{\int [p(x)f(x)+q(x)g(x)]dx}_{\text{primal=dual when reaching optimality}} && (2)\\
&= \underbrace{\iint [f(x)+g(y)]\gamma(x,y)dxdy}_{\text{marginal distribution}} && (3)\\
&\to f(x)+g(y)=c(x,y) && (4)
\end{aligned}$$

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

$$\begin{aligned}
&\underbrace{\int [p(x)f(x)+q(x)g(x)]dx}_{\text{dual formulation}} && (5)\\
&= \underbrace{\iint [f(x)+g(y)]\gamma(x,y)dxdy}_{\text{marginal distribution}} && (6)\\
&= \underbrace{\iint \gamma(x,y)c(x,y)dxdy}_{f(x)+g(y)=c(x,y)} && (7)\\
&\to \text{primal} = \text{dual when reaching optimality} && (8)
\end{aligned}$$

Lipschitz constrained formulation

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

1. consider the optimality condition when $x=y$
   1. $f(y)+g(y)=c(y,y)=0 \to g(y)=-f(y)$
2. take $g(y)=-f(y)$ into the $\mathcal{W}[p,q]$
   1. objective function: $\max_f\left\{\int [p(x)f(\boldsymbol{x})-q(\boldsymbol{x})f(\boldsymbol{x})]dx\right\}$
   2. constraints: $\|f\|_L\le 1$
      1. $f(x)-f(y)\le c(x,y)$ and $f(y)-f(x)\le c(y,x)$
      2. $\|f\|_L = \frac{|f(x)-f(y)|}{c(x,y)}\le 1$

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

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

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

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

    3. $f^{c\hat{c}}(x) \geq f(x)$, "$=$" 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

        $$\begin{aligned} & f(x) + \inf\{c(x,y) - f(x)\} && (9) \\ \leq & f(x) + c(x,y) - f(x) && (10) \\ = & c(x,y) && (11) \end{aligned}$$

        The constraint is always be satisfied under C-transform
        2. prove $f(x)$ and $f^c(y)$ reach optimality condition

        $$\begin{aligned} & f(x) = g^c(x) && (12) \\ \rightarrow & f^c(y) = g^{c\hat{c}}(y) \geq g(y) && (13) \\ \rightarrow & f(x) + f^c(y) \geq f(x) + g(y) && (14) \end{aligned}$$

        when $f(x) + g(y) = c(x,y)$, $c(x,y) \leq f(x) + f^c(y) \geq c(x,y)$
        Therefore $f(x) + f^c(y) = c(x,y)$ and reaches optimality.

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 $\gamma(x,y)$

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

Convex formulation

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

1. under quadratic case:

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

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:

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

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

$$\begin{aligned}&\mathcal{W}[p,q]=C_{p,q}-\min_{f'\in \mathrm{cvx}}\max_{g'\in \mathrm{cvx}}\left\{\mathbb{E}_p[f'(x)]+\mathbb{E}_q[f'^*(y)]\right\}\\&\mathcal{W}[p,q]=C_{p,q}-\min_{f'\in \mathrm{cvx}}\max_{g'\in \mathrm{cvx}}\left\{\mathbb{E}_p[f'(\nabla g'(y))]+\mathbb{E}_q[\langle\nabla g'(y),y\rangle-f'(\nabla g'(y))]\right\}\end{aligned}$$

Computation

Close form

1. Gaussian distribution under quadratic cost
    1. Distributions: $\mathcal{N}_1\left(\mu_1,\Sigma_1\right),\mathcal{N}_2\left(\mu_2,\Sigma_2\right)$
    2. Transport map: $x\longrightarrow\mu_2+A\left(x-\mu_1\right)$

1. $A=\Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)\Sigma_1^{-1/2}$
3. W-distance:

$$W_2\left(\mathcal{N}_1,\mathcal{N}_2\right)=\|\mu_1-\mu_2\|_2^2+\operatorname{Tr}\left(\Sigma_1+\Sigma_2-2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)$$

Discrete case

Linear programming

Maximize $\sum_{i=1}^n a_i f_i+\sum_{i=1}^n b_i g_i$
Subject to:

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

1. solve in dual form
2. polynomial complexity
3. not scalable when n is large

Sinkhorn iteration

Minimize $\sum_{i=1}^n\sum_{j=1}^n c_{ij}x_{ij}-\lambda^{-1}H(x)$
Subject to:

$$\begin{array}{l}\sum_{j=1}^n x_{ij}=p_i\quad \text{for } i=1,2,\ldots,n\\ \sum_{i=1}^n x_{ij}=q_j\quad \text{for } j=1,2,\ldots,n\\ x_{ij}\geq 0\quad \text{for } i,j=1,2,\ldots,n\end{array}$$

1. entropy regularization: $H(x)=-x\log x$
   1. strongly convex
   2. link primal and dual solution
2. Optimality condition
   1. $\nabla_x L(x,f,g)=0=c_{ij}+\lambda^{-1}(1+\log x)-f_i-g_j$
   2. $x_{ij}=e^{\lambda f_i}e^{-c_{ij}\lambda-1}e^{\lambda g_j}=v_iK_{ij}u_j$
   3. constraints

$$\begin{array}{l}\sum_{j=1}^n x_{ij}=v_i\sum_{j=1}^n K_{ij}u_j=p_i\quad \text{for } i=1,2,\ldots,n\\ \sum_{i=1}^n x_{ij}=u_j\sum_{i=1}^n v_iK_{ij}=q_j\quad \text{for } j=1,2,\ldots,n\end{array}$$

3. Matrix normalization/balancing
   1. find a matrix has row and column constraints
   2. double stochastic matrix
4. Sinkhorn-Knopp algorithm

$$\begin{aligned}v_i^{n+1}&=\frac{p_i}{\sum_j K_{ij}u_j^n}\\ u_j^{n+1}&=\frac{p_i}{\sum_i K_{ij}v_i^{n+1}}\end{aligned}$$

5. limited numerical accuracy when λ is large

ADMM

Continue case

Two-step computation

Penalty term

Lipschitz constraint

Unconstrained optimization

Convex formulation

Gaussian mixture model

Empirical study

Application