Optimal Transport And WGAN

Yiping Lu[1]

[1]School of Mathmetical Science
Peking University

Data Science Seminar, PKU, 2017

Outline

1. Introduction To Optimal Transport.
2. Minkowski Type Problems
   - Picewise Linear Function And Power Diagram
3. WGAN

Outline

Optimal Transport

Monge

Objective: Calculate a transport map $T_{\#}\mu = \upsilon$ which minimize the transport cost

$$c(T) = \int c(x, T(x)) d\mu(x)$$

Kantorovich

Objective: Calculate a transport plane minimize the transport cost

$$c(\Pi) = \int c(x, y)\Pi(x, y)$$

Optimal Transport: Linear Programming View

The optimal transport is a convex problem, which can be formulated as

$$\begin{aligned} &\min \langle C, F\rangle \\ s.t. &\sum_i F_{i,j} = q_j \\ &\sum_j F_{i,j} = q_i \end{aligned}$$

is a special case of the linear programming:

$$\begin{aligned} &\min c^T x \\ &s.t. Ax = b, x \geq 0 \end{aligned}$$

Linear Programming

Consider the dual problem of the linear programming problem.

Primal:

$$\min c^T x$$
$$s.t. Ax = b, x \geq 0$$

Dual:

$$\min b^T y$$
$$s.t. A^T y \leq c$$

and we have the relation:

$$\begin{aligned}
\inf_{Ax=b, x\geq 0} c^T x &= \inf_{x\geq 0} \sup_{y} c^T x + y^T(b - Ax) \\
&=^{?} \sup_{y} \inf_{x \geq 0} c^T x - y^T A x + y^T b \\
&= \sup_{A^T y \leq c} y^T b
\end{aligned}$$

Kantorovich Dual

First, let us express the constraint $\gamma \in \Pi(\mu, \upsilon)$ in the following way.

$$\sup_{\phi,\psi} \int_X \phi d\mu + \int_Y \psi d\upsilon - \int_{X\times Y} (\phi(x) + \psi(y)) d\gamma$$

so that the primal problem can be expressed by

$$\min_{\gamma \geq 0} \int_{X\times Y} + \sup_{\phi,\psi} \int_X \phi d\mu + \int_Y \psi d\upsilon - \int_{X\times Y} (\phi(x) + \psi(y)) d\gamma$$

then consider interchanging sup and inf:

$$\sup_{\phi,\psi} \int_X \phi d\mu + \int_Y \psi d\upsilon + \inf_{\gamma \geq 0} \int_{X\times Y} (c(x,y) - (\phi(x) + \psi(y))) d\gamma$$

Kantorovich Dual

If the central notion inthe original Monge-Kantorovich problem is cost, in the dual problem it is price.

Imagine that a company offers to take care of all your transportation problem, buying bread at the bakeries and selling them to the cafes.

Let $\psi(x)$ be the price at which a basker of bread at the bakery x and selling them to the cafe y at the price $\phi(y)$

Let us maximize the profit:

$$\sup \left\{ \int_Y \phi(y) d\upsilon(y) - \int_X \psi(x) d\mu(y) | \phi(y) - \psi(x) \leq c(x,y) \right\}$$

Kantorovich Dual

It is easy to proof that

$$\sup_{\phi-\psi\leq c}\left\{\int_Y \phi(y)d\upsilon(y) - \int_X \psi(x)d\mu(y)\right\}$$
$$\leq \inf_{\pi\in\Pi(\mu,\upsilon)}\left\{\int_{X\times Y} c(x,y)d\pi(x,y)\right\}$$

If we describe a pair of prices (ϕ,ψ) as tight if

$$\phi(y) = \inf_x(\psi(x) + c(x,y))$$
$$\psi(x) = \sup_y(\phi(y) - c(x,y))$$

The following formula can be seen as the definition of $c-$transform.

c-Cyclical Monotonicity

Definition

Once a function $c : X\times \to \mathbb{R} \cup \{+\infty\}$ is given , we say that a set $\Gamma \subset X \times Y$ is c-cyclically monotone if for every $k \in \mathbb{N}$, every permutation σ and every finite family of points $(x_1, y_1), \cdots, (x_k, y_k) \in \Gamma$ we have

$$\sum_{i=1}^{k} c(x_i, y_i) \leq \sum_{i=1}^{k} c(x_i, y_{\sigma(i)})$$

c-Cyclical Monotonicity

Definition

Once a function $c : X\times \to \mathbb{R} \cup \{+\infty\}$ is given , we say that a set $\Gamma \subset X \times Y$ is c-cyclically monotone if for every $k \in \mathbb{N}$, every permutation σ and every finite family of points $(x_1, y_1), \cdots, (x_k, y_k) \in \Gamma$ we have

$$\sum_{i=1}^{k} c(x_i, y_i) \leq \sum_{i=1}^{k} c(x_i, y_{\sigma(i)})$$

Theorem

If γ is an optimal transport plane for the cost c and c is continuous, then spt(γ) is a CM-set.

c-Cyclical Monotonicity

Theorem

Rockafellar's Theorem

If $\Gamma \neq \emptyset$ is a $c-$CM set in $X \times Y$ and $c : X \times Y \to \mathbb{R}$, then there exists a $c-$concave function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ such that

$$\Gamma \subset \{(x, y) \in X \times Y : \phi(x) + \phi^c(y) = c(x, y)\}$$

c-Cyclical Monotonicity

Theorem

Rockafellar's Theorem

If $\Gamma \neq \emptyset$ is a $c-$CM set in $X \times Y$ and $c : X \times Y \to \mathbb{R}$, then there exists a $c-$concave function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ such that

$$\Gamma \subset \{(x,y) \in X \times Y : \phi(x) + \phi^c(y) = c(x,y)\}$$

Proof.

The function ϕ can be defined as

$$\begin{aligned}\phi(x) = \inf\{c(x,y_n) - c(x_n,y_n) + c(x_n,y_{n-1}) - c(x_{n-1},y_{n-1}) + \cdots \\ +c(x_1,y_0) - c(x_0,y_0) : n \in \mathbb{N}, (x_i,y_i) \in \Gamma\}\end{aligned}$$

□

Kantorovich Dual

As a result, we can get a theorem as below.

Theorem

If c is C^1, ϕ is a Kantorovich potential for the cost c in the transport from μ to υ, and (x_0, y_0) belongs to the support of an optimal transport plane γ, then $\nabla_\phi(x_0) = \nabla_x c(x_0, y_0)$, provided ϕ is differentiable at x_0.

Kantorovich Dual

As a result, we can get a theorem as below.

Theorem

If c is C^1, ϕ is a Kantorovich potential for the cost c in the transport from μ to υ, and (x_0, y_0) belongs to the support of an optimal transport plane γ, then $\nabla_\phi(x_0) = \nabla_x c(x_0, y_0)$, provided ϕ is differentiable at x_0.

As an example, if the cost function has the **following form** $c(x, y) = h(x - y)$**, h is strictly convex**. Then there is exists an optimal transport plan γ for the cost $c(x, y)$ and is unique of the form $(id, T)_{\#}\mu$.
Moreover, ther exists a Kantorovich potential ϕ and T and the potentials ϕ are linked by

$$T(x) = x - (\nabla h)^{-1}(\nabla \phi(x))$$

Quadratic Case

For the quadratic case $c(x,y)=\frac{1}{2}|x-y|^2$

$$T(x)=x-\nabla\phi(x)=\nabla(\frac{x^2}{2}-\phi(x))=\nabla u(x)$$

Theorem

For function $X:\mathbb{R}^n\to\mathbb{R}$, let us define $u_X=\frac{1}{2}|x|^2-X(x)$, then we have

$$u_{X^c}=(u_X)^*$$

Proof.

$$u_{X^c}(x)=\sup_y\frac{1}{2}|x|^2-\frac{1}{2}|x-y|^2+X(y)=\sup_y x\cdot y-\left(\frac{1}{2}|y|^2-X(y)\right)$$

$\square$

Quadratic Case

We go futhur more on the quadratic case, we only need to minimize the $\int x \cdot y d\gamma$ gives the same result.
We can give the same result easier, actually we have $\phi(x_0) + \phi^*(y_0) = x_0 \cdot y_0$ for $y_0 \in \partial\phi(x_0)$, which means

Theorem

For the quadratic case, there exists unique an optimal transport map T from μ to υ and it is of the form $T = \nabla u$ for a convex function u

Quadratic Case

We go futhur more on the quadratic case, we only need to minimize the $\int x \cdot y d\gamma$ gives the same result.
We can give the same result easier, actually we have $\phi(x_0) + \phi^*(y_0) = x_0 \cdot y_0$ for $y_0 \in \partial\phi(x_0)$, which means

Theorem

For the quadratic case, there exists unique an optimal transport map T from μ to υ and it is of the form $T = \nabla u$ for a convex function u

Remark

- The ϕ above is called **Kantorovich Potential**.
- The u here is called **Brenier Potential**.

Outline

Minkowski Problem

Theorem

Suppose Ω is a compact convex polytope with non-empty interior in $\mathbb{R}^n$ are distinct k points and $A_1, A_2, \cdots, A_k > 0$ s.t. $\sum_{i=1}^{k} A_i = vol(\Omega)$. Then there exists a vector $h = (h_1, \cdots, h_k) \in \mathbb{R}^k$, unique up to adding the constant $(c, c, \cdots, c)$, so that the piecewise linear convex function

$$u(x) = \max_{x \in \Omega} \{x \cdot p_i + h_i\}$$

satisfies $vol(\{x \in \Omega | \nabla u(x) = p_i\}) = A_i$

Minkowski Problem

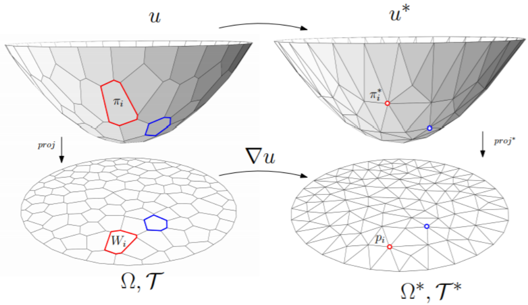


Figure 2: Discrete Optimal Transport Mapping (left to right): map W_i to p_i. Discrete Monge-Ampere equation (right to left): $vol(W_i)$ is the discrete Hessian determinant of p_i.

Outline

1. Introduction To Optimal Transport.
2. Minkowski Type Problems
   - Picewise Linear Function And Power Diagram
3. WGAN

Piecewise Linear(PL) Function

Definition

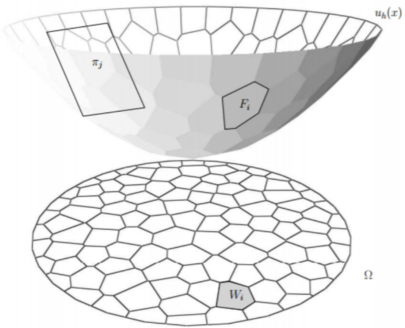
PL Function: For $P = \{p_1, \cdots, p_k\}$ and $h = (h_1, \cdots, h_k) \in \mathbb{R}^k$, we define the PL convex function $u_h(x)$ to be

$$u(x) = \max\{p_i \cdot x + h_i | i = 1, 2, \cdots, k\}$$

The domain $D(u^*)$ of the dual function u^* is the convex hull of P and

$$u^*(y) = \min\{-\sum_{i=1}^{k} t_i h_i | t_i \geq 0, \sum_{i=1}^{k} t_i = 1, \sum_{i=1}^{k} t_i p_i = y\}$$

Piecewise Linear(PL) Function



PL-convex function f defined on a closed convex polyhedron produces a convex subdivision.

Power Diagram

Definition

(Power Distance) Given a point $y_i \in \mathbb{R}^n$ with a power weight ϕ_i the power distance is given by

$$pow(x, y_i) = |x - y_i|^2 - \phi_i$$

Definition

(Power Diagram) Given weighted points $\{(y_i, \phi_i)\}$, the power diagram is the cell decompostion of $\mathbb{R}^n$, denote as $V(\phi)$

$$\mathbb{R}^n = \cup_{i=1}^k W_i(\phi), W_i(\phi) = \{x \in \mathbb{R}^n | pow(x, y_i) \leq pow(x, y_j), \forall j\}$$

Each cell is a convex polytope

Power Diagram

Now consider a eqaul construction as let $h_i = \frac{1}{2}(\phi_i - |y_i|^2)$, we construct the convex function

$$u_h(x) = \max_i\{\langle x, y_i\rangle + h\}, W_i(h) = \max_i\{x \cdot p_i + h_i\}$$

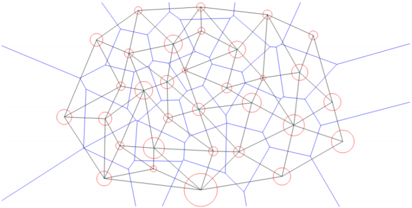


Figure 5: Power diagram (blue) and its dual weighted Delaunay triangulation (black), the power weight ψ_i equal to the square of radius r_i (red circle).

Variation

Proposition

Suppose $\sigma \to \mathbb{R}$ is continuous defined on a compact convex domain $\Omega \subset \mathbb{R}^n$. If $p_1, \cdots, p_k \in \mathbb{R}^n$ are distinct and $h \in \mathbb{R}^k$ so that $vol(W_i(h) \cap \Omega) > 0$ for all i, then $\omega_i(h) = \int_{W_i(h)\cap\Omega} \sigma(x)$ is a differentialable function in h so that for $j \neq i$ and $W_i(h) \cap \Omega$ and $W_i(h) \cap \Omega$ share a codimension-1 face F,

$$\frac{\partial \omega_i(h)}{\partial h_j} = -\frac{1}{|p_i - p_j|} \int_F \sigma_F(x) dA$$

where dA is the area form on F and parital derivative is zero otherwise.

Variation

It is easy to observe that $\frac{\partial \omega_i}{\partial h_j} = \frac{\partial \omega_j}{\partial h_i}$, thus we can give our main theorem.

Theorem

Theorem 4.3 (Gu-Luo-Sun-Yau[12]) *Let Ω be a compact convex domain in $\mathbb{R}^n$, $\{y_1, ..., y_k\}$ be a set of distinct points in $\mathbb{R}^n$ and μ a probability measure on Ω. Then for any $\nu_1, ..., \nu_k > 0$ with $\sum_{i=1}^k \nu_i = \mu(\Omega)$, there exists $h = (h_1, ..., h_k) \in \mathbb{R}^k$, unique up to adding a constant $(c, ..., c)$, so that $w_i(h) = \nu_i$, for all i. The vectors h are exactly maximum points of the concave function*

$$E(h) = \sum_{i=1}^k h_i \nu_i - \int_0^h \sum_{i=1}^k w_i(\eta) d\eta_i \tag{19}$$

on the open convex set

$$H = \{h \in \mathbb{R}^k | w_i(h) > 0, \forall i\}.$$

Furthermore, ∇u_h minimizes the quadratic cost

$$\int_\Omega |x - T(x)|^2 d\mu(x)$$

among all transport maps $T_\# \mu = \nu$, where the Dirac measure $\nu = \sum_{i=1}^k \nu_i \delta_{y_i}$.

Semi-discrete Optimal Mass Transport

For our empirical distribution is defined as the sum of several Dirac measure $v = \sum_{j=1}^{k} v_j \delta(y - y_j)$
Define the discrete Kantorovich potential $\phi : Y \to \mathbb{R}, \phi(y_j) = \phi_j$, then

$$\int_Y \phi dv = \sum_{j=1}^{k} \phi_j v_j$$

Define the $c-$transformation of ϕ is given by

$$\phi^c(x) = \min_{1 \leq j \leq k} \{c(x, y_j) - \phi_j\}$$

and each cell is defined as

$$W_i(\phi) = \{x \in X | c(x, y_i) - \phi_i \leq c(x, y_j) - \phi_j, \forall 1 \leq j \leq k\}$$

Brenier's Approach

We only consider the situation that the cost function is the L^2 distance. Here

$$u_h(x) = \max_{i=1}^{k}\{\langle x, y_i\rangle + h\}$$

Then

$$W_i(h) = \{x \in X | \nabla u_h(x) = y_i\} \cap \Omega$$

and at the same time

$$\nabla u_h : W_i(h) \to y_i, i = 1, 2, \cdots, k$$

Outline

GAN/WGAN

Objective Function:

$$\min_{u \in U} \max_{v} \mathbb{E}_{x \sim D_{real}}[\phi(D_v(x))] + \mathbb{E}_{x \sim D_G}[\phi(1 - D_v(x))]$$

- GAN:$\phi = \log$
- WGAN:$\phi = \mathbf{id}$

GAN/WGAN

Objective Function:

$$\min_{u\in U}\max_{v}\mathbb{E}_{x\sim D_{real}}[\phi(D_v(x))]+\mathbb{E}_{x\sim D_G}[\phi(1-D_v(x))]$$

- GAN: $\phi = \log$
- WGAN: $\phi = \mathbf{id}$

Rewrite the WGAN Objective:

$$\min_{u\in U}\max_{v}\mathbb{E}_{x\sim D_{real}}[D_v(x)]+\mathbb{E}_{x\sim D_G}[1-D_v(x)]$$

GAN/WGAN

Objective Function:

$$\min_{u \in U} \max_{v} \mathbb{E}_{x \sim D_{real}}[\phi(D_v(x))] + \mathbb{E}_{x \sim D_G}[\phi(1 - D_v(x))]$$

- GAN:$\phi = \log$
- WGAN:$\phi = \mathbf{id}$

Rewrite the WGAN Objective:

$$\min_{u \in U} \max_{v} \mathbb{E}_{x \sim D_{real}}[D_v(x)] + \mathbb{E}_{x \sim D_G}[1 - D_v(x)]$$

equal to

$$\min_{u \in U} \max_{v} \mathbb{E}_{x \sim D_{real}}[D_v(x)] - \mathbb{E}_{x \sim D_G}[D_v(x)]$$

GAN/WGAN

Objective Function:

$$\min_{u\in U}\max_{v}\mathbb{E}_{x\sim D_{real}}[\phi(D_v(x))]+\mathbb{E}_{x\sim D_G}[\phi(1-D_v(x))]$$

- GAN:$\phi=\log$
- WGAN:$\phi=$ **id**

Rewrite the WGAN Objective:

$$\min_{u\in U}\max_{v}\mathbb{E}_{x\sim D_{real}}[D_v(x)]+\mathbb{E}_{x\sim D_G}[1-D_v(x)]$$

equal to

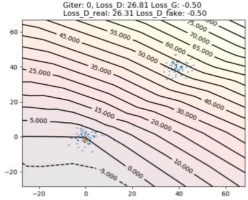
$$\min_{u\in U}\max_{v}\mathbb{E}_{x\sim D_{real}}[D_v(x)]-\mathbb{E}_{x\sim D_G}[D_v(x)]$$

For if $c(x,y)=|x-y|$, then $\phi^c=-\phi$(1-Lip),**approximate to W_1 Distance**

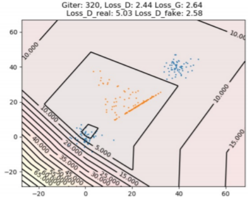
Geometric Generative Model

- Ecoding/Decoding process: This setp maps the smaples between the image space X and the latent space Z by deep neural networks, the encoding map is denoted as $f_\theta : X \to Z$ and decoding map is $g_\xi : Z \to X$
- Probability measure transformation process: This step transform a fixed distibution $\xi \in P(Z)$ to any given distribution $\mu \in P(Z)$, the mapping is denoted as $T : Z \to Z$, $T_{\#}\xi = \mu$. This step can either use conventional deep neural network or use explicit geometric/numerical methods.

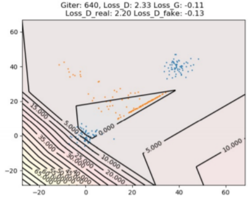
WGAN



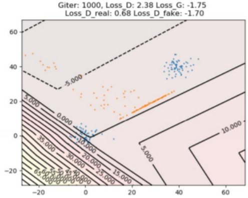
(a) initial stage



(b) after 320 iterations

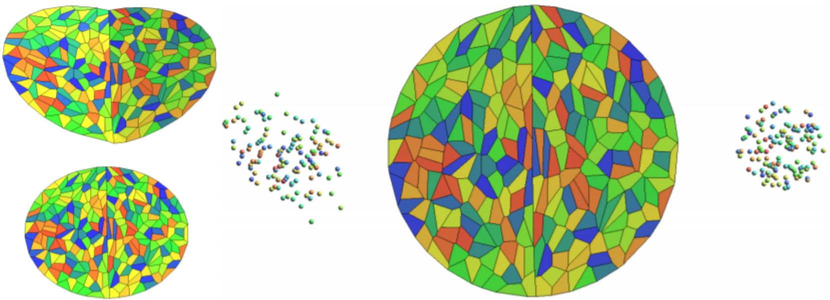


(c) after 640 iterations



(d) final stage, after 1000 iterations

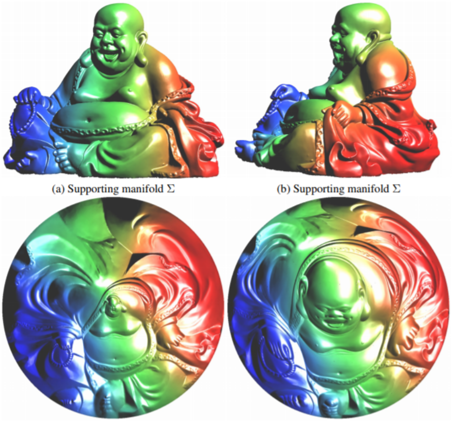
Geometric OMT



(a) Brenier potential u_h (b) Optimal transportation map T: power diagram cell $W_i \mapsto y_i$

Figure 9: Geometric model learns the Gaussian mixture distribution .

Geometric Method



(a) Supporting manifold Σ

(b) Supporting manifold Σ

(c) Image of encoding map $f_\theta(\Sigma)$

(d) Image of encoding map composed with optimal transportation map $(T^{-1} \circ f_\theta)(\Sigma)$

Figure 10: Illustration of geometric generative model.

Geometric Method

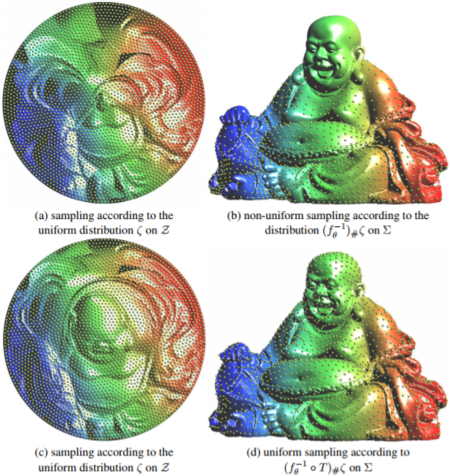


Figure 11: Illustration of geometric generative model.

Conclusion

1. 生成器：最优映射等价于Power胞腔分解，将每个胞腔W_i映到y_i，
2. 判别器：Wasserstein距离中$W_c(\mu,\nu)$中的ψ等于power 权重，
3. 判别器：Wasserstein距离Kantorovich势能φ等于power距离，
   $\varphi(x) = \min_i\{\text{pow}(x, y_i)\}$
4. 生成器：Brenier势能等于Power Diagram的上包络。