

# Optimal Transportation: Duality Theory

David Gu

Computer Science Department  
Stony Brook University

*gu@cs.stonybrook.edu*

September 1, 2022

# Motivation

# Why dose DL work?

## Problem

- ① *What does a DL system really learn ?*
- ② *How does a DL system learn ? Does it really learn or just memorize ?*
- ③ *How well does a DL system learn ? Does it really learn everything or have to forget something ?*

Till today, the understanding of deep learning remains primitive.

# Why does DL work?

1. What does a DL system really learn?

*Probability distributions on manifolds.*

2. How does a DL system learn ? Does it really learn or just memorize ?

*Optimization in the space of all probability distributions on a manifold. A DL system both learns and memorizes.*

3. How well does a DL system learn ? Does it really learn everything or have to forget something ?

*Current DL systems have fundamental flaws, mode collapsing.*

# Manifold Distribution Principle

We believe the great success of deep learning can be partially explained by the well accepted manifold distribution and the clustering distribution principles:

## Manifold Distribution

A natural data class can be treated as a probability distribution defined on a low dimensional manifold embedded in a high dimensional ambient space.

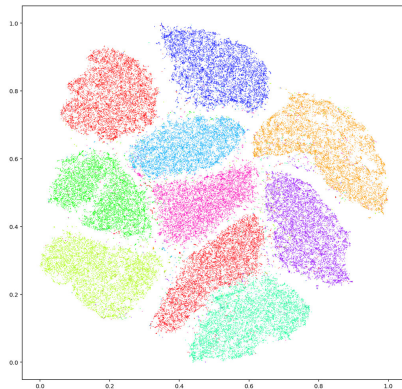
## Clustering Distribution

The distances among the probability distributions of subclasses on the manifold are far enough to discriminate them.

# MNIST tSNE Embedding



a. LeCunn's MNIST

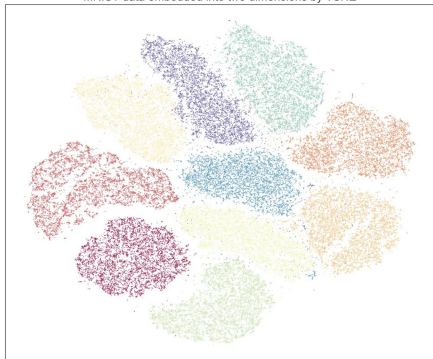


b. Hinton's t-SNE embedding

- Each image  $28 \times 28$  is treated as a point in the image space  $\mathbb{R}^{28 \times 28}$ ;
- The hand-written digits image manifold is only two dimensional;
- Each digit corresponds to a distribution on the manifold.

# Manifold Learning

MNIST data embedded into two dimensions by TSNE



MNIST data embedded into two dimensions by UMAP

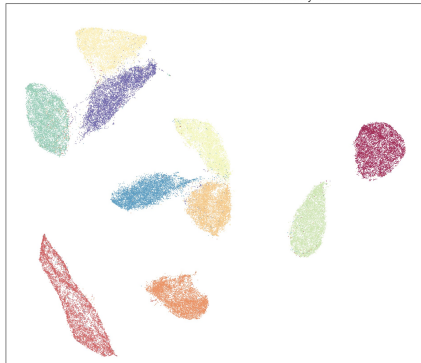


Figure: t-SNE embedding and UMap embedding.

# How does a DL system learn ?

## Optimization

- Given a manifold  $X$ , all the probability distributions on  $X$  form an infinite dimensional manifold, Wasserstein Space  $\mathcal{P}(X)$ ;
- Deep Learning tasks are reduced to optimization in  $\mathcal{P}(X)$ , such as the principle of maximum entropy principle, maximum likely hood estimation, maximum a posterior estimation and so on;
- DL tasks requires variational calculus, Riemannian metric structure defined on  $\mathcal{P}(X)$ .



## Solution

- Optimal transport theory discovers a natural Riemannian metric of  $\mathcal{P}(X)$ , called Wasserstein metric;
- the covariant calculus on  $\mathcal{P}(X)$  can be defined accordingly;
- the optimization in  $\mathcal{P}(X)$  can be carried out.

# Equivalence to Conventional DL Methods

- Entropy function is convex along the geodesics on  $\mathcal{P}(X)$ ;
- The Hessian of entropy defines another Riemannian metric of  $\mathcal{P}(X)$ ;
- The Wasserstein metric and the Hessian metric are equivalent in general;
- Entropy optimization is the foundation of Deep Learning;
- Therefore Wasserstein-metric driven optimization is equivalent to entropy optimization.

# Optimal Transportation

- The geodesic distance between  $d\mu = f(x)dx$  and  $d\nu(y) = g(y)dy$  is given by the optimal transport map  $T : X \rightarrow X$ ,  $T = \nabla u$ ,

$$\det \left( \frac{\partial^2 u}{\partial x_i \partial x_j} \right) = \frac{f(x)}{g \circ \nabla u(x)}.$$

- The geodesic between them is McCann's displacement,

$$\gamma(t) := ((1-t)I + t\nabla u)_{\#}\mu$$

- The tangent vectors of a probability measure is a gradient field on  $X$ , the Riemannian metric is given by

$$\langle d\varphi_1, d\varphi_2 \rangle = \int_X \langle d\varphi_1, d\varphi_2 \rangle_{\mathbf{g}} f(x) dx.$$

# How well does a DL system learn ?

Fundamental flaws: mode collapsing and mode mixture.

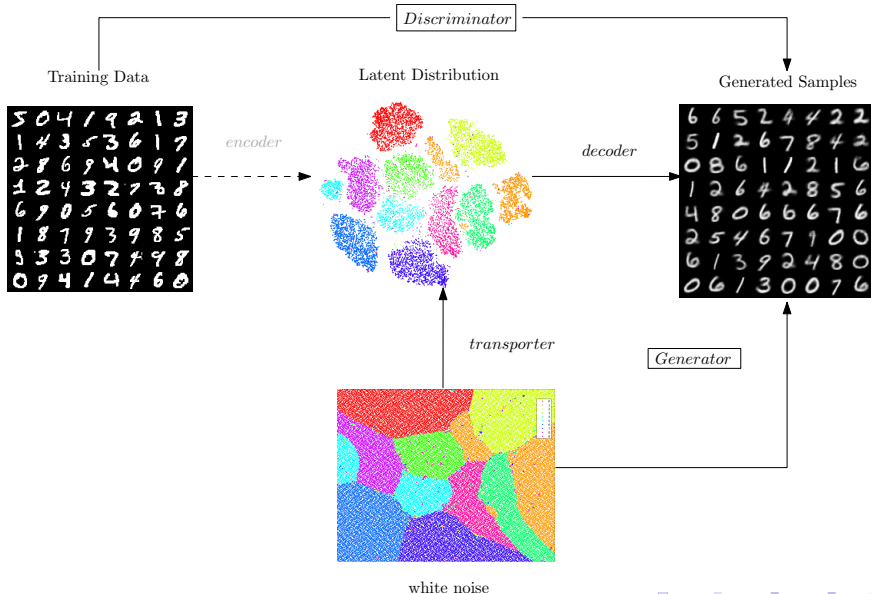


(a). VAE

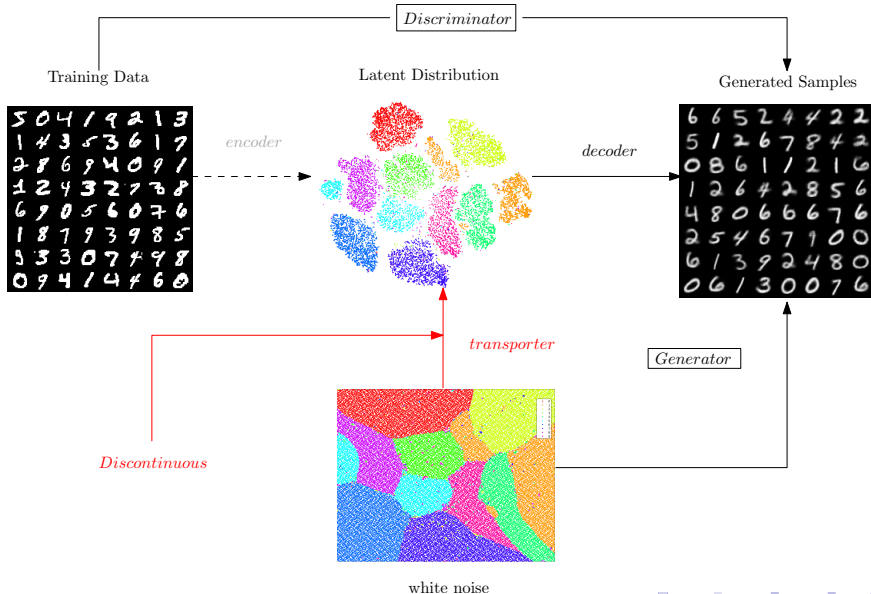


(b). WGAN

# GAN model



# GAN model - Mode Collapse Reason



# Mode Collapse Reason

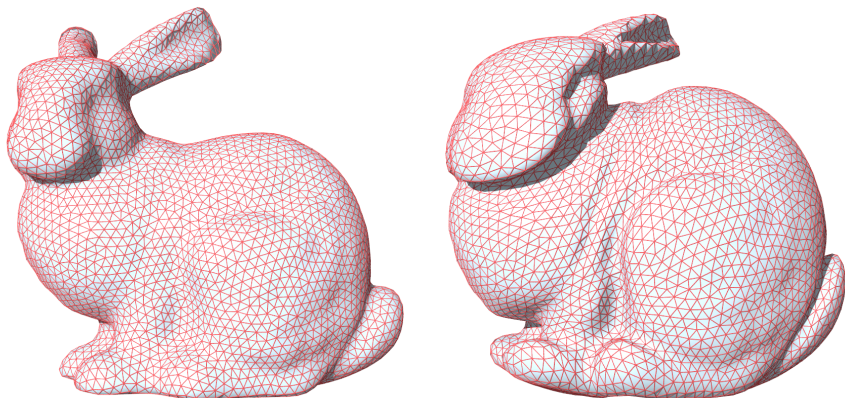


Figure: Singularities of an OT map.

# Mode Collapse Reason

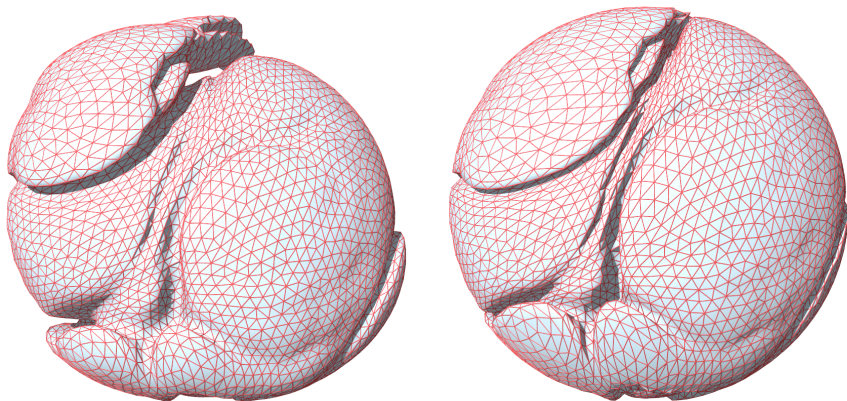


Figure: Singularities of an OT map.



# How to eliminate mode collapse?

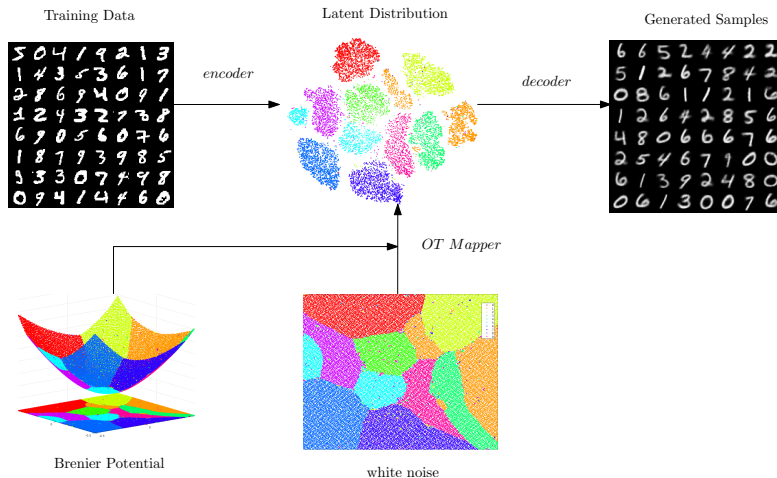
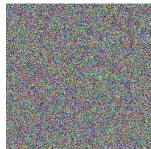


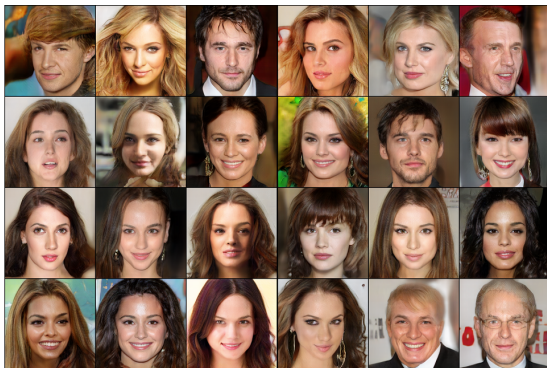
Figure: Geometric Generative Model.

# Generative and Adversarial Networks

Noise  $\sim N(0,1)$



Generative  
Model



A generative model converts a white noise into a facial image.

# Generative and Adversarial Networks



A GAN model based on OT theory.

There are three views of optimal transportation theory:

- 1 Duality view
- 2 Fluid dynamics view
- 3 Differential geometric view

Different views give different insights and induce different computational methods; but all three theories are coherent and consistent.

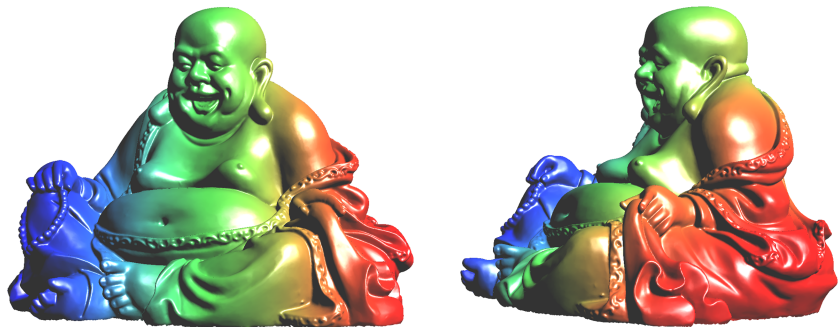


Figure: Buddha surface.

# Optimal Transportation Map

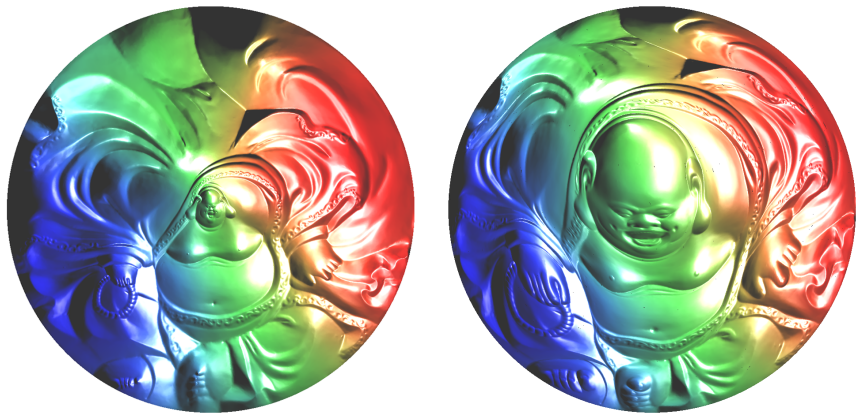


Figure: Optimal transportation map.

# Optimal Transportation Map

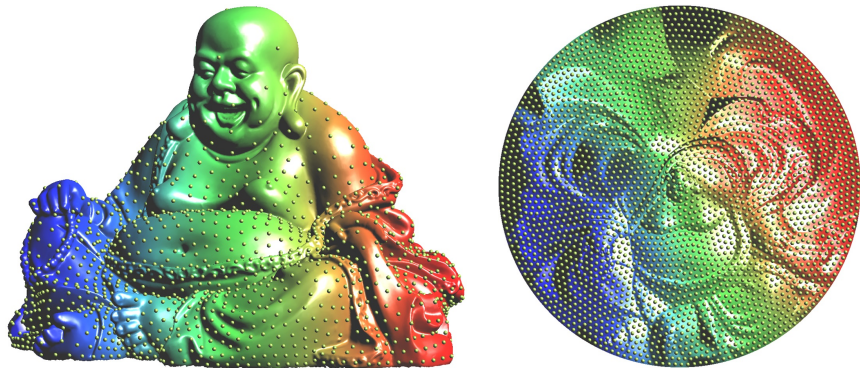


Figure: Brenier potential.

# Optimal Transportation Map

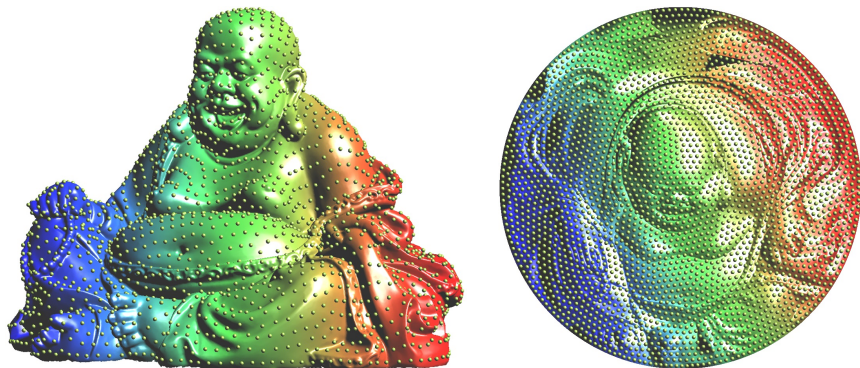


Figure: Brenier potential.



# Optimal Transportation Map

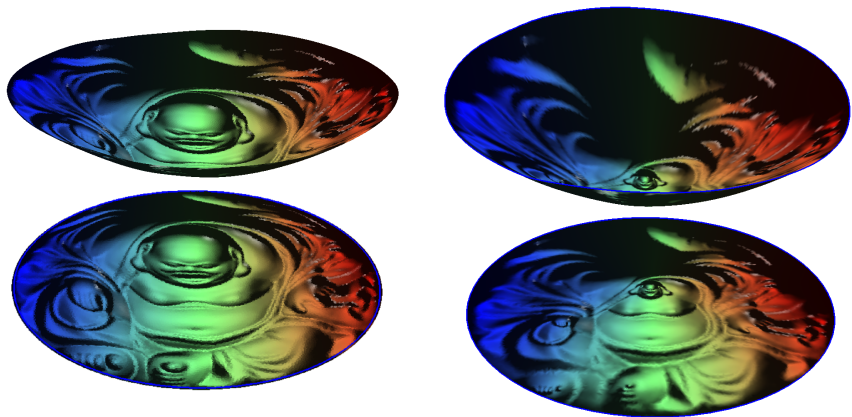


Figure: Brenier potential.

# Duality Theories

# Monge Problem

Assume  $\Omega$  and  $\Sigma$  are two domains in the Euclidean space,  $\mathbb{R}^d$ ,  $\mu$  and  $\nu$  are two probability measures on  $\Omega$  and  $\Sigma$  respectively,  $\mu \in \mathcal{P}(\Omega)$ ,  $\nu \in \mathcal{P}(\Sigma)$ , such that they have equal total measure:

$$\mu(\Omega) = \nu(\Sigma). \quad (1)$$

## Definition (Measure-preserving Map)

A mapping  $T : \Omega \rightarrow \Sigma$  is called *measure preserving*, if for any Borel set  $B \subset \Sigma$ ,

$$\int_{T^{-1}(B)} d\mu = \int_B d\nu, \quad (2)$$

and is denoted as  $T_{\#}\mu = \nu$   $T$  pushes  $\mu$  forward to  $\nu$ .

# Monge Problem

Suppose the density functions of  $\mu$  and  $\nu$  are given by  $f : \Omega \rightarrow \mathbb{R}$  and  $g : \Sigma \rightarrow \mathbb{R}$ , namely

$$d\mu = f(x_1, x_2, \dots, x_d) dx_1 \wedge dx_2 \wedge \dots \wedge dx_d,$$

$$d\nu = g(y_1, y_2, \dots, y_d) dy_1 \wedge dy_2 \wedge \dots \wedge dy_d,$$

and  $T : \Omega \rightarrow \Sigma$  is  $C^1$  and measure-preserving,

$$f(x_1, \dots, x_d) dx_1 \wedge \dots \wedge dx_d = g(T(x)) dy_1 \wedge \dots \wedge dy_d.$$

then  $T$  satisfies the Jacobi equation:

## Definition (Jacobi Equation)

$$\det DT(x) = \frac{f(x)}{g \circ T(x)} \quad (3)$$

# Monge Problem

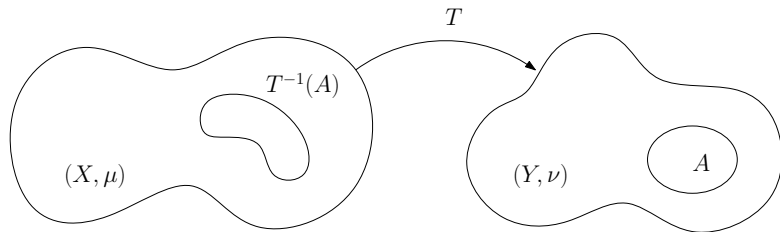


Figure: Measure-preserving map.

# Monge Problem

## Definition (Transportation Cost)

Given a cost function  $c : \Omega \times \Sigma \rightarrow \mathbb{R}$ , the total transportation cost for a map  $T : \Omega \rightarrow \Sigma$  is defined as

$$\mathcal{C}(T) := \int_{\Omega} c(x, T(x)) d\mu(x).$$

## Problem (Monge)

Among all the measure-preserving mappings,  $T : \Omega \rightarrow \Sigma$  and  $T_{\#}\mu = \nu$ , find the one with the minimal total transportation cost,

$$MP : \quad \min \left\{ \int_{\Omega} c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\}. \quad (4)$$

# Monge Problem

## Definition (Optimal Transportation Map)

The solution to the Monge problem is called an optimal transportation map between  $(\Omega, \mu)$  and  $(\Sigma, \nu)$ .

Suppose  $\Omega$  coincides with  $\Sigma$

## Definition (Wasserstein Distance)

The total cost of the optimal transportation map  $T : \Omega \rightarrow \Sigma$ ,  $T_{\#}\mu = \nu$ , is called the Wasserstein distance between  $\mu$  and  $\nu$ .

Suppose the cost is the square of the Euclidean distance  $c(x, y) = |x - y|^2$ , then the Wasserstein distance is defined as

$$\mathcal{W}_2^2(\mu, \nu) := \inf \left\{ \int_{\Omega} |x - T(x)|^2 d\mu(x) : T_{\#}\mu = \nu \right\}.$$

# Kantorovich Problem

## Transportation Plan

Kantorovich relax the transportation map to transportation scheme, or transportation plan, which is represented by a joint probability distribution  $\rho : \omega \times \Sigma \rightarrow \mathbb{R}$ ,  $\rho(x, y)$  represents how much mass is transported from the source point  $x$  to the target point  $y$ .

## Marginal Distribution

The marginal distribution of  $\rho$  equals to  $\mu$  and  $\nu$ , namely we have the condition

$$(\pi_x)_\# \rho = \mu, \quad (\pi_y)_\# \rho = \nu, \quad (5)$$

where the projection maps

$$\pi_x(x, y) = x, \quad \pi_y(x, y) = y.$$



## Transportation map vs. Transportation plan

Transportation map is a special case of transportation plan, namely a transportation map  $T : \Omega \rightarrow \Sigma$  induces a transportation plan

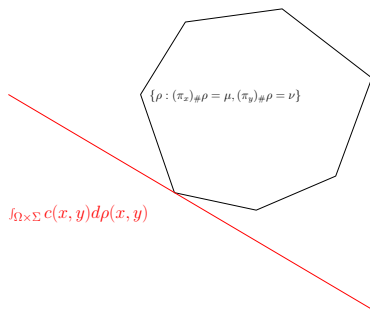
$$(Id, T)_{\#}\mu = \rho. \quad (6)$$

# Kantorovich Problem

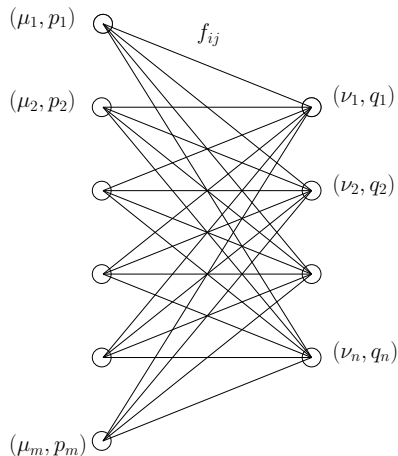
## Problem (Kantorovich )

Find a transportation plan with the minimal total transportation cost,

$$KP : \min \left\{ \int_{\Omega \times \Sigma} c(x, y) d\rho(x, y) : (\pi_x)_\# \rho = \mu, (\pi_y)_\# \rho = \nu \right\}. \quad (7)$$



# Kantorovich Problem



## Problem (Linear Programming)

$$\min \sum_{ij} c(p_i, q_j) f_{ij},$$

such that

$$\forall i, \sum_j f_{ij} = \mu_i$$

$$\forall j, \sum_i f_{ij} = \nu_j.$$

# Kantorovich Problem

## Linear Programming

Kantorovich problem is to find a minimal value of a linear function defined on a convex polytope, so the solution exists. KP can be solved using linear programming method, such as simplex, interior point or ellipsoid algorithms.

## Kantorovich Problem

In general situation, the support of a transportation plan  $\rho$  covers all the  $\Omega \times \Sigma$ . If the transportation map  $T$  exists, the support of  $(Id, T)_\# \mu$  has 0 measure in  $\Omega \times \Sigma$ . KP doesn't discover the intrinsic structure, it is highly inefficient to compute optimal transportation map.

# Kantorovich Problem

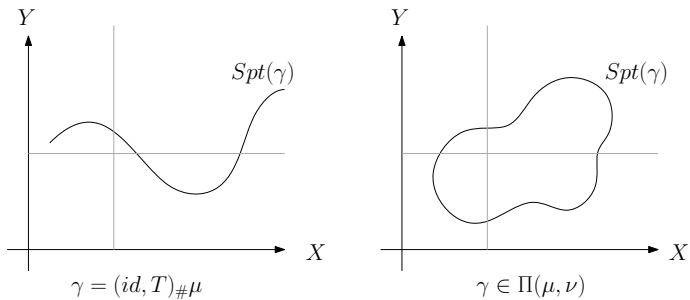


Figure: Caption

# Kantorovich Dual Problem

Denote  $\Pi(\mu, \nu) = \{\rho : (\pi_x)_\# \rho = \mu, (\pi_y)_\# \rho = \nu\}$ . We consider the constraint  $\rho \in \Pi(\mu, \nu)$ . we have

$$\sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu - \int_{\Omega \times \Sigma} (\varphi(x) + \psi(y)) d\rho = \begin{cases} 0 & \rho \in \Pi(\mu, \nu), \\ +\infty & \rho \notin \Pi(\mu, \nu), \end{cases} \quad (8)$$

where the superimum is taken among all bounded continuous functions,  $\varphi \in C_b(\Omega)$  and  $\psi \in C_b(\Sigma)$ .

# Kantorovich Dual Problem

We use this as a generalized Lagrange multiplier in (KP), and rewrite (KP) as

$$\min_{\rho} \int_{\Omega \times \Sigma} c d\rho + \sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu - \int_{\Omega \times \Sigma} (\varphi(x) + \psi(y)) d\rho \quad (9)$$

Under suitable conditions, such as Rockafella's conditions, we can exchange sup and inf

$$\sup_{\varphi, \psi} \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu + \inf_{\rho} \int_{\Omega \times \Sigma} (c(x, y) - (\varphi(x) + \psi(y))) d\rho. \quad (10)$$

We can rewrite the infimum in  $\rho$  as a constraint on  $\varphi$  and  $\psi$ :

$$\inf_{\rho \geq 0} \int_{\Omega \times \Sigma} (c - \varphi \oplus \psi) d\rho = \begin{cases} 0 & \varphi \oplus \psi \leq c \text{ on } X \times Y \\ -\infty & \varphi \oplus \psi > c \end{cases}$$

where  $\varphi \oplus \psi$  denotes the function  $\varphi \oplus \psi(x, y) := \varphi(x) + \psi(y)$ .

# Kantorovich Dual Problem

This leads to the dual optimization problem.

## Problem (Dual)

Given  $\mu \in \mathcal{P}(\Omega)$  and  $\nu \in \mathcal{P}(\Sigma)$  and the cost function  $c : \Omega \times \Sigma \rightarrow [0, +\infty)$ , we consider the problem

$$(DP) \quad \max \left\{ \int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu : \varphi \in C_b(\Omega), \psi \in C_b(\Sigma) : \varphi \oplus \psi \leq c \right\}. \quad (11)$$

From the condition  $\varphi \oplus \psi \leq c$ , we obtain  $\sup DP \leq \min KP$ ,

$$\int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu = \int_{\Omega \times \Sigma} \varphi \oplus \psi d\rho \leq \int_{\Omega \times \Sigma} c d\rho$$

This is valid for all admissible pairs  $(\varphi, \psi)$  and every admissible  $\rho$ .



# Kantorovich Dual Problem

From the condition  $\varphi \oplus \psi \leq c$ , we obtain  $\sup DP \leq \min KP$ ,

$$\int_{\Omega} \varphi d\mu + \int_{\Sigma} \psi d\nu = \int_{\Omega \times \Sigma} \varphi \oplus \psi d\rho \leq \int_{\Omega \times \Sigma} c d\rho$$

This is valid for all admissible pairs  $(\varphi, \psi)$  and every admissible  $\rho$ . This shows

$$\boxed{\max(DP) \leq \min(KP)}$$

## Definition (c-transform)

Given  $\varphi \in L^1(\Omega)$ , and the cost function  $c : \Omega \times \Sigma \rightarrow \mathbb{R}$ , the c-transform of  $\varphi$  is defined as  $\varphi^c : \Sigma \rightarrow \mathbb{R}$ ,

$$\varphi^c(y) := \inf_{x \in \Omega} c(x, y) - \varphi(x), \quad (12)$$

The optimization of Kantorovich functional is equivalent to replace the Kantorovich potentials  $(\varphi_n, \psi_n)$  by the c-transforms of the other, namely

$$(\varphi, \psi) \rightarrow (\varphi, \varphi^c) \rightarrow (\varphi^{cc}, \varphi^c) \rightarrow (\varphi^{cc}, \varphi^{ccc}) \dots$$

# c-transform

Geometrically, if we fix a point  $x \in \Omega$ , then we get a supporting surface

$$\Gamma_x : \Sigma \rightarrow \mathbb{R},$$

$$\Gamma_x(y) := c(x, y) - \varphi(x),$$

the graph of the c-transform  $\varphi^c(y)$  is the envelope of all these supporting surfaces.

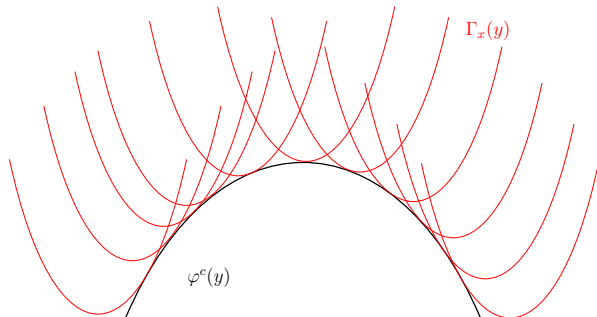


Figure: Geometric interpretation of c-transform.

# Twisting Condition

By  $\varphi^c(y) = \inf_x c(x, y) - \varphi(x)$ , we obtain

$$\nabla_x c(x, y(x)) = \nabla \varphi(x)$$

## Definition (Twisting condition)

Given a cost function  $c : \Omega \times \Sigma \rightarrow \mathbb{R}$ , if for any  $x \in \Omega$ , the mapping

$$\mathcal{L}_x(y) := \nabla_x c(x, y)$$

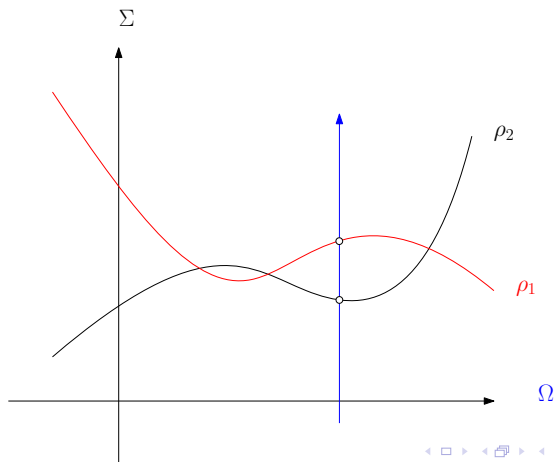
is injective, then we say  $c$  satisfies twisting condition.

If  $c$  satisfies the twisting condition, then an optimal plan is an optimal map.

# Uniqueness of Optimal Transportation Map

## Theorem (Uniqueness)

*Suppose  $c$  satisfies the twisting condition, then the optimal transportation map is unique.*



# Uniqueness of Optimal Transportation Map

## Proof.

Assume there are two optimal transportation maps  $T_1, T_2 : (\Omega, \mu) \rightarrow (\Sigma, \nu)$ , the corresponding optimal transportation plans are

$$\rho_k = (Id, T_k)_\# \mu, \quad k = 1, 2.$$

Then  $\frac{1}{2}(\rho_1 + \rho_2)$  is also an optimal transportation. Since  $c$  satisfies the twisting condition,  $\frac{1}{2}(\rho_1 + \rho_2)$  corresponds to an optimal transport map. But the blue line intersects the support of  $\frac{1}{2}(\rho_1 + \rho_2)$  at two points, it is not a map. Contradiction. □

By utilizing c-transform, we obtain

## Problem (Dual Problem)

Given  $\mu \in \mathcal{P}(\Omega)$ ,  $\nu \in \mathcal{P}(\Sigma)$ , the dual problem is

$$DP : \max_{\varphi \in C_b(\Omega)} \left\{ \int_{\Omega} \varphi(x) d\mu(x) + \int_{\Sigma} \varphi^c(y) d\nu(y) \right\}. \quad (13)$$

# Cyclic Monotonicity

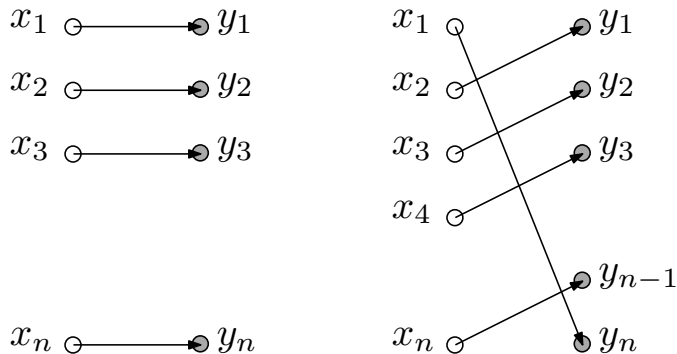


Figure: Cyclic monotonicity.

$\rho$  is optimal, then for any  $(x, y) \in \text{Supp}(\rho)$ ,  $\varphi(x) + \psi(y) = c(x, y)$ .



## Definition (Cyclic Monotonicity)

Suppose  $\Gamma \subset \mathbb{R}^d$  is a domain, for any set of pair of points:

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k) \subset \text{Supp}(\rho),$$

we have the following inequality

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

where  $\sigma$  is a permutation of  $1, 2, \dots, k$ , then we say  $\Gamma$  is cyclic monotonous.

The cyclic monotonicity can be applied to prove the equivalence between Kantorovich problem and Kantorovic dual problem.

# Cyclic Monotonicity

## Definition (c-concave)

A function  $\varphi : \Omega \rightarrow \mathbb{R}$  is called c-concave, if there is a function  $\psi : \Omega \rightarrow \mathbb{R}$ , such that  $\varphi = \psi^c$ .

## Theorem

If  $\Gamma \neq \emptyset$ ,  $\Gamma$  is cyclic monotonous in  $\Omega \times \Sigma$ , then there exists a c-concave function  $\varphi$ , such that

$$\Gamma \subset \{(x, y) \in \Omega \times \Sigma : \varphi(x) + \varphi^c(y) = c(x, y)\}.$$

## Theorem

If  $\rho$  is an optimal transport plan for the continuous cost  $c$ , then its support  $\text{supp}(\rho)$  is cyclic monotonous.

# Cyclic Monotonicity

## Theorem ( $\max(DP) = \min(KP)$ )

Suppose that  $\Omega$  and  $\Sigma$  are Polish spaces and that  $c : \Omega \times \Sigma \rightarrow \mathbb{R}$  is uniformly continuous and bounded. Then the problem (DP) admits a solution  $(\varphi, \varphi^c)$  and we have

$$\max(DP) = \min(KP)$$

## Proof.

Suppose  $\rho$  is a solution to (KP), then  $\text{Supp}(\rho)$  satisfies cyclic monotonicity; hence there exists  $\varphi$  and  $\varphi^c$ ,  $\text{Supp}(\rho) \subset \{\varphi + \varphi^c = c\}$ , therefore

$$\min(KP) = \int_{\Omega \times \Sigma} c d\rho \leq \int_{\Omega} \varphi d\mu + \int_{\Sigma} \varphi^c d\nu \leq \max(DP).$$

# Monge-Ampere Equation

## Lemma

Suppose  $c : \Omega \rightarrow \mathbb{R}$  is a  $C^2$  strictly convex function,  $\Omega$  is convex, then  $\nabla c : \Omega \rightarrow \mathbb{R}^d$  is injective.

## Proof.

Suppose there are two distinct points  $x_0, x_1 \in \Omega$ , such that  $\nabla c(x_0) = \nabla c(x_1)$ . Draw a line segment  $\gamma : [0, 1] \rightarrow \Omega$ ,  $\gamma(0) = x_0$  and  $\gamma(1) = x_1$ . Then  $f(t) = c \circ \gamma(t)$  is strictly convex

$$f'(t) = \langle \nabla c((1-t)x_0 + tx_1), x_1 - x_0 \rangle$$

$$f''(t) = (x_1 - x_0)^T D^2 c((1-t)x_0 + tx_1) (x_1 - x_0).$$

Therefore,  $f'(1) = f'(0)$  and  $f''(t) > 0$ . Contradiction. □

# Monge-Ampere Equation

## Lemma

Suppose  $c : \Omega \rightarrow \mathbb{R}$  is a strictly convex function,  $\Omega$  is convex, then  $\nabla c : \Omega \rightarrow \mathbb{R}^d$  is injective.

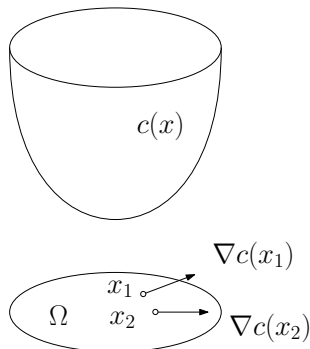


Figure: Injectivity of the gradient map of a strictly convex function.

# Monge-Ampere Equation

Suppose the cost function is a strictly convex function, satisfying the condition  $c(x, y) = c(x - y)$ , then

$$D_x c(x, y) - D\varphi(x) = 0,$$

we obtain  $D_x c(x - y) = D\varphi(x)$ ,

$$T(x) = y = x - (Dc)^{-1}(D\varphi(x)),$$

# Brenier Problem

## Theorem (Brenier)

Given  $\mu \in \mathcal{P}(\Omega)$  and  $\nu \in \mathcal{P}(\Sigma)$ , and the cost function  $c(x, y) = \frac{1}{2}|x - y|^2$ , the optimal transportation map is the gradient of a function  $u : \Omega \rightarrow \mathbb{R}$ ,  $T(x) := \nabla u(x)$ .

## Proof.

We obtain

$$T(x) = x - D\varphi(x) = D\left(\frac{|x|^2}{2} - \varphi(x)\right) = Du(x).$$



# Brenier Problem

## Problem (Brenier)

Find a convex function  $u : \Omega \rightarrow \mathbb{R}$ , satisfying the Monge-Ampère equation,

$$\det \left( \frac{\partial^2 u(x)}{\partial x_i \partial x_j} \right) = \frac{f(x)}{g \circ \nabla u(x)}. \quad (14)$$

## Proof.

We plug  $T(x) = Du(x)$  into the Jacobi equation, we obtain the Monge-Ampère equation,

$$\det DT = \frac{f(x)}{g \circ T(x)}$$

hence

$$\det \left( \frac{\partial^2 u(x)}{\partial x_i \partial x_j} \right) = \frac{f(x)}{g \circ \nabla u(x)}.$$