Optimal Transportation: Duality Theory

David Gu

Yau Mathematics Science Center
Tsinghua University
Computer Science Department
Stony Brook University

gu@cs.stonybrook.edu

August 15, 2020

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

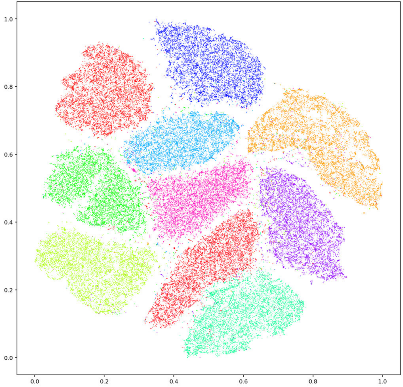
A natural data class can be treated as a probability distribution defined on a low dimensional manifold embedded in a high dimensional ambient space.

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 embemdding

- Each image 28×28 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

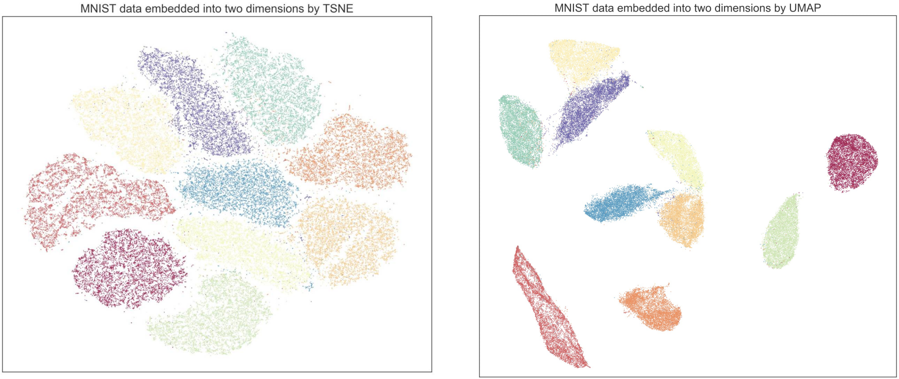


Figure: t-SNE embedding and UMap embedding.

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 $T : X \to X$, $T = \nabla u$,

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

- The geodesic between them is McCann's displacement,

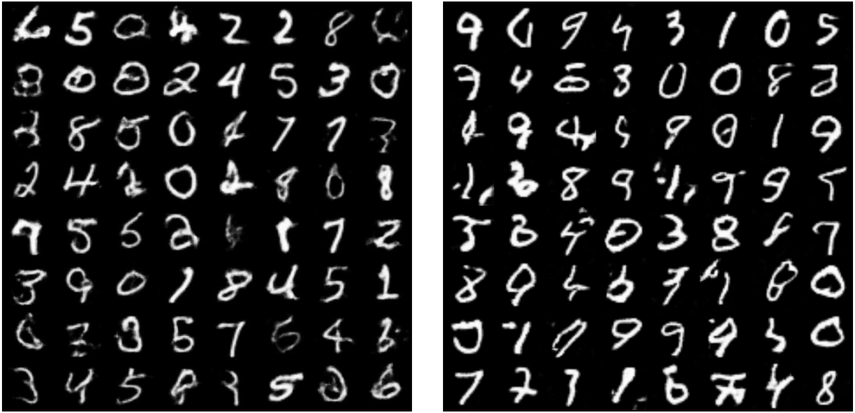
$$\gamma(t) := ((1-t)I + t\nabla u)_{\#}\mu$$

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

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

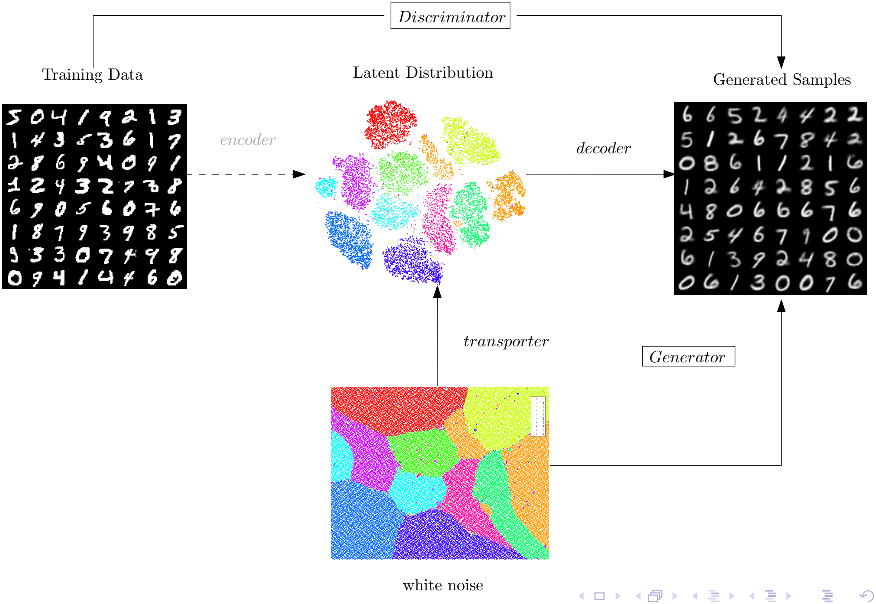
How well does a DL system learn ?

Fundamental flaws: mode collapsing and mode mixture.



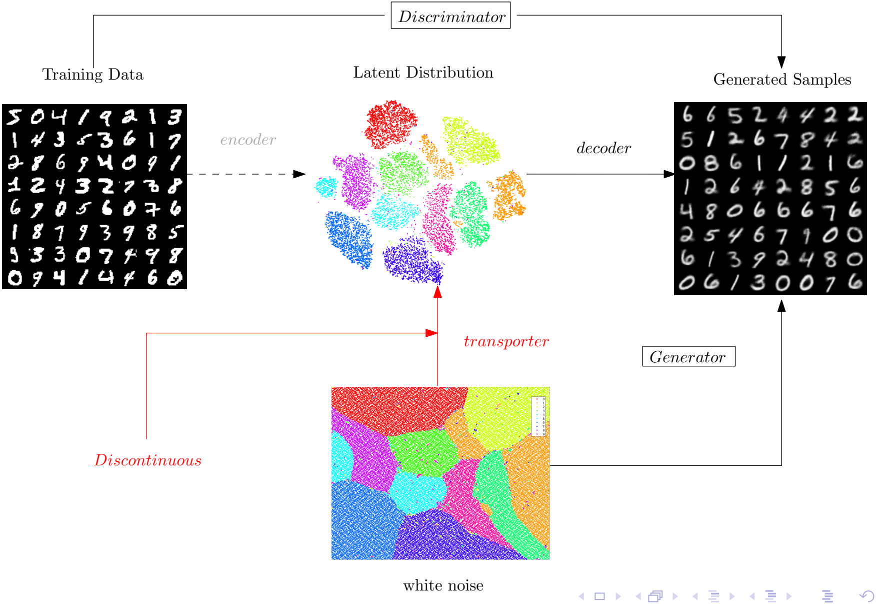
(a). VAE (b). WGAN

GAN model



Discriminator
Training Data
Latent Distribution
Generated Samples
encoder
decoder
transporter
Generator
white noise

GAN model - Mode Collapse Reason



Mode Collapse Reason

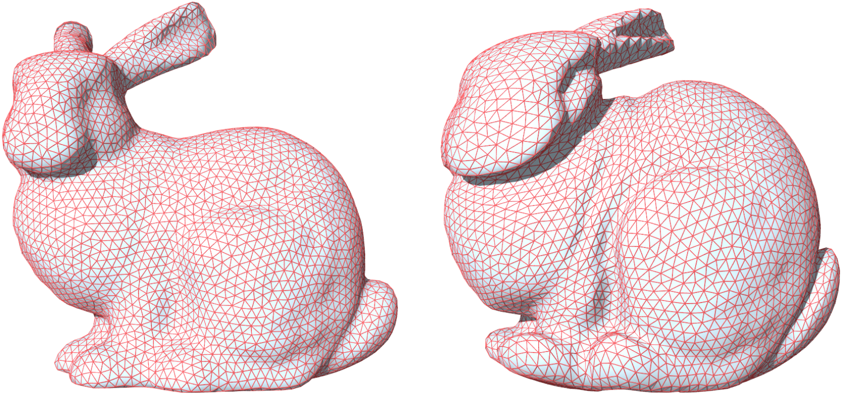


Figure: Singularities of an OT map.

Mode Collapse Reason

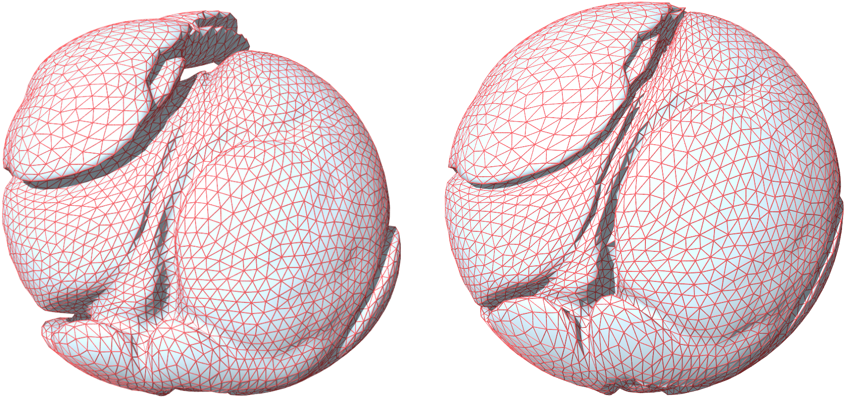


Figure: Singularities of an OT map.

How to eliminate mode collapse?

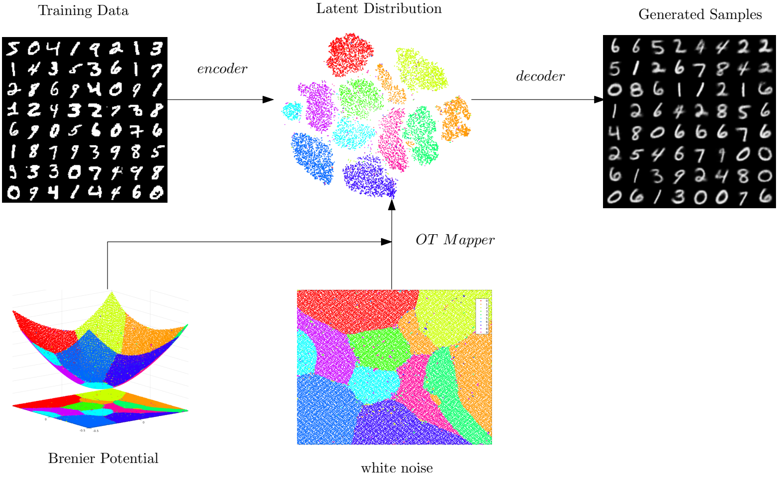
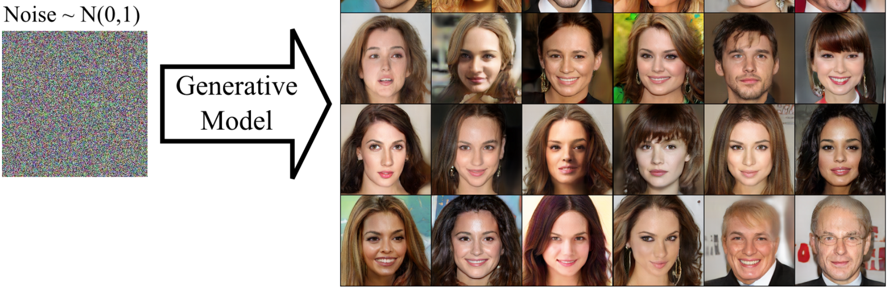


Figure: Geometric Generative Model.

Generative and Adverse rial Networks



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

Generative and Adverse rial Networks

A GAN model based on OT theory.

Overview

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.

Optimal Transportation Map

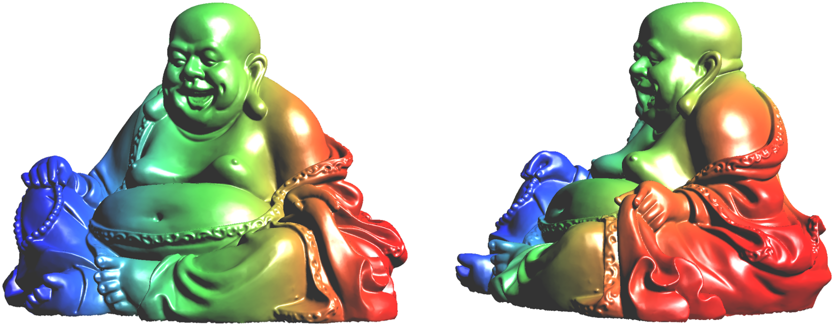


Figure: Buddha surface.

Optimal Transportation Map

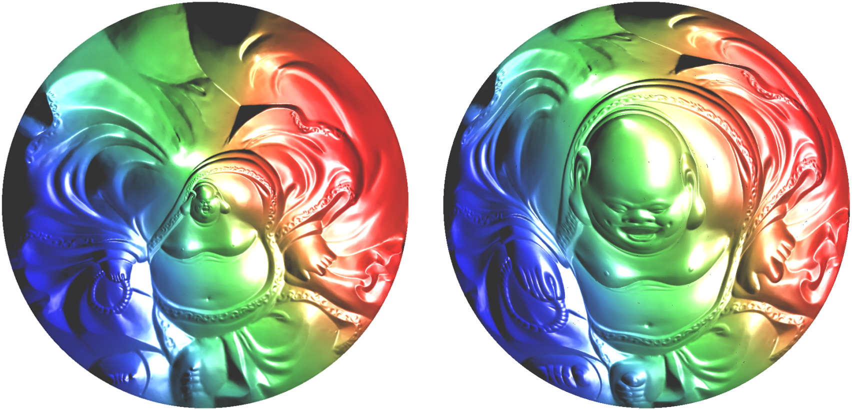


Figure: Optimal transportation map.

Optimal Transportation Map

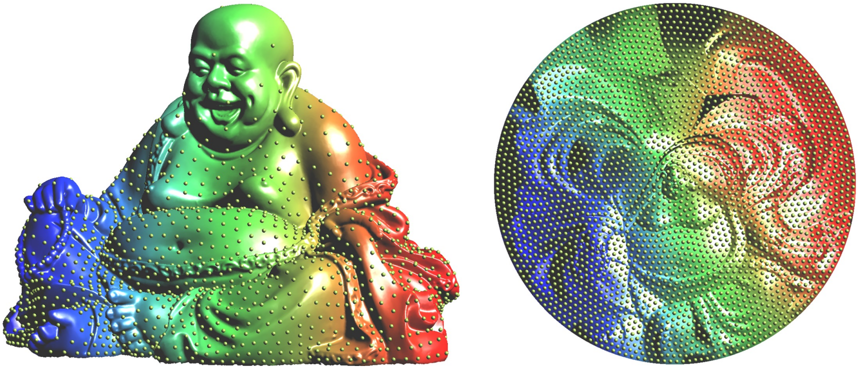


Figure: Brenier potential.

Optimal Transportation Map

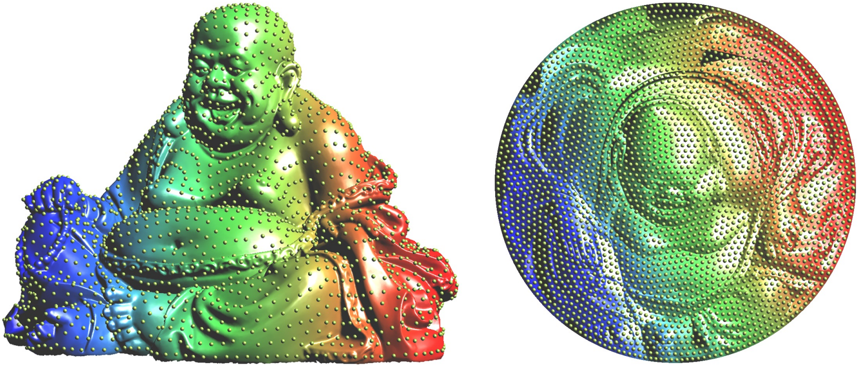


Figure: Brenier potential.

Optimal Transportation Map

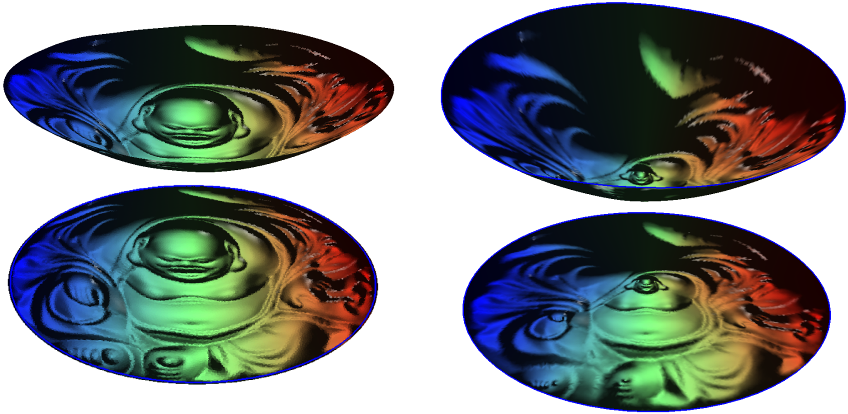


Figure: Brenier potential.

Duality Theories

Monge Problem

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

$$\mu(\Omega) = \nu(\Sigma). \tag{1}$$

Definition (Measure-preserving Map)

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

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

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

Monge Problem

Suppose the density functions of μ and ν are given by $f : \Omega \to \mathbb{R}$ and $g : \Sigma \to \mathbb{R}$, namely

$$d\mu = f(x_1, x_2, \cdots, x_d) dx_1 \wedge dx_2 \wedge \cdots \wedge dx_d,$$
$$d\nu = g(y_1, y_2, \cdots, y_d) dy_1 \wedge dy_2 \wedge \cdots \wedge dy_d,$$

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

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

then T satisfies the Jacobi equation:

Definition (Jacobi Equation)

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

Monge Problem

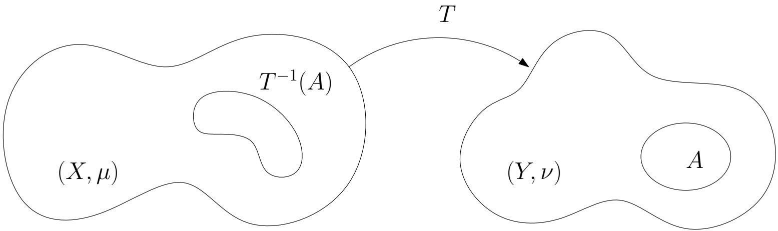


Figure: Measure-preserving map.

Monge Problem

Definition (Transportation Cost)

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

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

Problem (Monge)

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

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

Monge Problem

Definition (Optimal Transportation Map)

The solution to the Monge problem is called an optimal transportation map between (Ω, μ) and (Σ, ν).

Suppose Ω coincides with Σ

Definition (Wasserstein Distance)

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

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

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

Kantorovich Problem

Transportation Plan

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

Marginal Distribution

The marginal districution of ρ equals to μ and ν, namely we have the condition

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

where the projection maps

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

Kantorovich Problem

Transportation map vs. Transportation plan

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

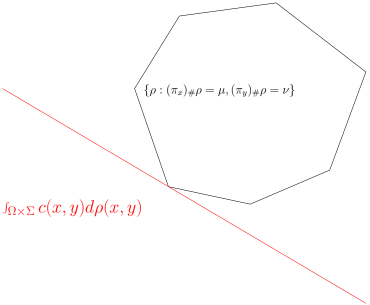
$$(Id, T)_{\#}\mu = \rho. \tag{6}$$

Kantorovich Problem

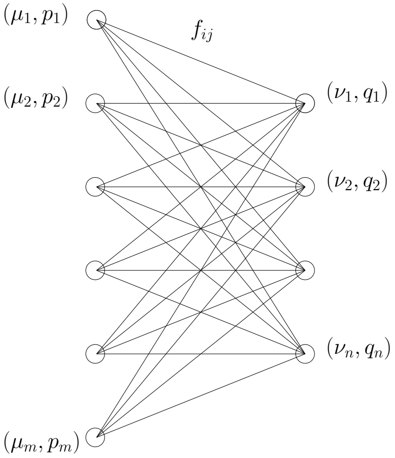
Problem (Kantorovich)

Find a transportation plan with the minimal toal transportation cost,

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



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 simplx, interior point or ellipsoid algorithms.

Kantorovich Problem

In general situation, the support of a transportation plan ρ 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

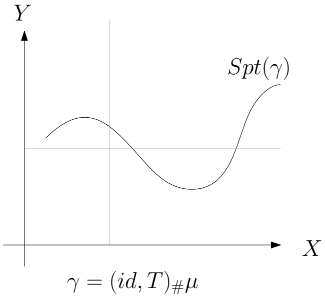


Figure: Caption

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

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

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

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

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

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

We can rewrite the infimum in ρ as a constraint on φ and ψ:

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

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

Kantovorich 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* $c : \Omega \times \Sigma \to [0, +\infty)$, *we consider the problem*

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

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 (φ, ψ) and every admissible ρ.

Kantovorich 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 (φ, ψ) and every admissible ρ. This shows

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

Definition (c-transform)

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

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

The optimization of Kantorovich functional is equivalent to replace the Kantorovich potentials (φ_n, ψ_n) by the c-transforms of the other, namely

$$(\varphi, \psi) \to (\varphi, \varphi^c) \to (\varphi^{cc}, \varphi^c) \to (\varphi^{cc}, \varphi^{ccc}) \cdots$$

c-transform

Geometrically, if we fix a point $x \in \Omega$, then we get a supporting surface $\Gamma_x : \Sigma \to \mathbb{R}$,

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

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

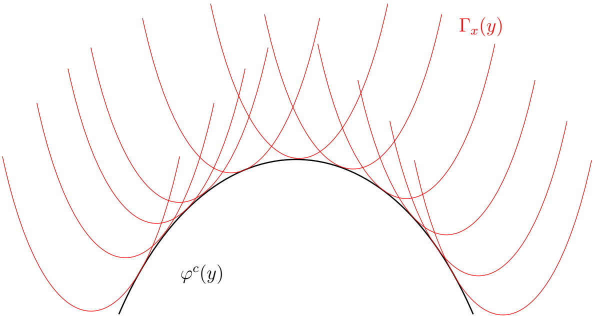


Figure: Geometric interpretation of c-transform.

Twisting Condition

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

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

Definition (Twisting condition)

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

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

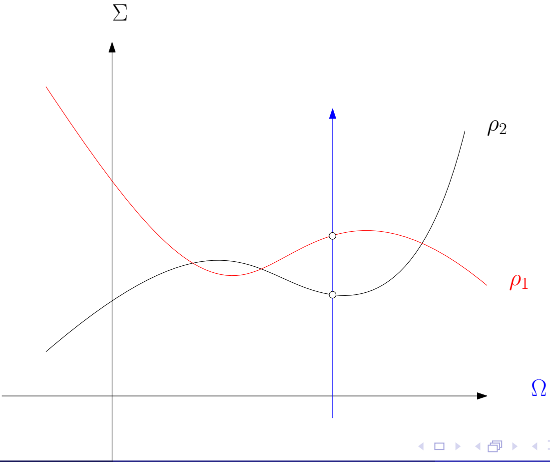
is injective, then we say c satisfies twisting condition.

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 transportion maps $T_1, T_2 : (\Omega, \mu) \to (\Sigma, \nu)$, 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. $\square$

Dual Problem

By utilizing c-transform, we obtain

Problem (Dual Problem)

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

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

Cyclic Monotonocity

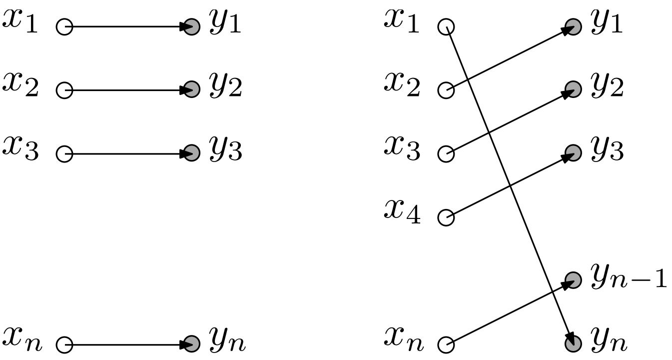


Figure: Cyclic monotonocity.

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

Cyclic Monotonocity

Definition (Cyclic Monotonocity)

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

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

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 σ is a permuation of $1, 2, \ldots, k$, then we say Γ is cyclic monotonous.

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

Cyclic Monotonocity

Definition (c-concave)

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

Theorem

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

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

Theorem

If ρ is an optimal transport plan for the continuous cost c, then its support supp(ρ) is cyclic monotonous.

Cyclic Monotonocity

Theorem (max (DP)=min (KP))

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

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

Proof.

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

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

Lemma

Suppose $c : \Omega \to \mathbb{R}$ is a C^2 strictly convex function, Ω is convex, then $\nabla c : \Omega \to \mathbb{R}^d$ 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 $\gamma : [0,1] \to \Omega$, $\gamma(0) = x_0$ 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. $\square$

Monge-Ampere Equation

Lemma

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

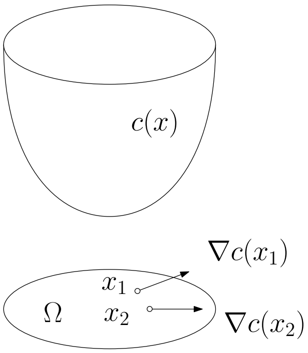


Figure: Injectivity of the gradient map of a strictly convex function.

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 $u : \Omega \to \mathbb{R}$, $T(x) := \nabla u(x)$.

Proof.

We obtain

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

$\square$

Brenier Problem

Problem (Brenier)

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

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

Proof.

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

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

hence

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

□