Geometric Understanding of Deep Learning

David Xianfeng Gu[1]

[1]Computer Science Department
Applied Mathematics Department
Stonybrook University

AI Institute
Stony Brook University

Thanks

Thanks for the invitation.

Collaborators

These projects are collaborated with Prof. Shing-Tung Yau, Prof. Dimitris Samaras, Prof. Feng Luo and many other mathematicians, computer scientists.

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

Helmholtz Hypothesis

Helmholtz Hypothesis

Half of the brain is devoted to vision, Helmholtz hypothesized that vision solves an inverse problem, i.e., inferring the most likely causes of the retina image. In modern language, the brain learns a generative model of visual images, and visual perception is to infer the latent variables of this generative model. The generative model with its multiple layers of latent variables form a representation of our visual world.

About representation learning, the basic idea is that the brain represents a concept by a group of neurons, or latent variables, that form a vector. Sometimes it is also called embedding, i.e., we embed the concept into a multi-dimensional Euclidean space, which is sometimes called latent space.

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

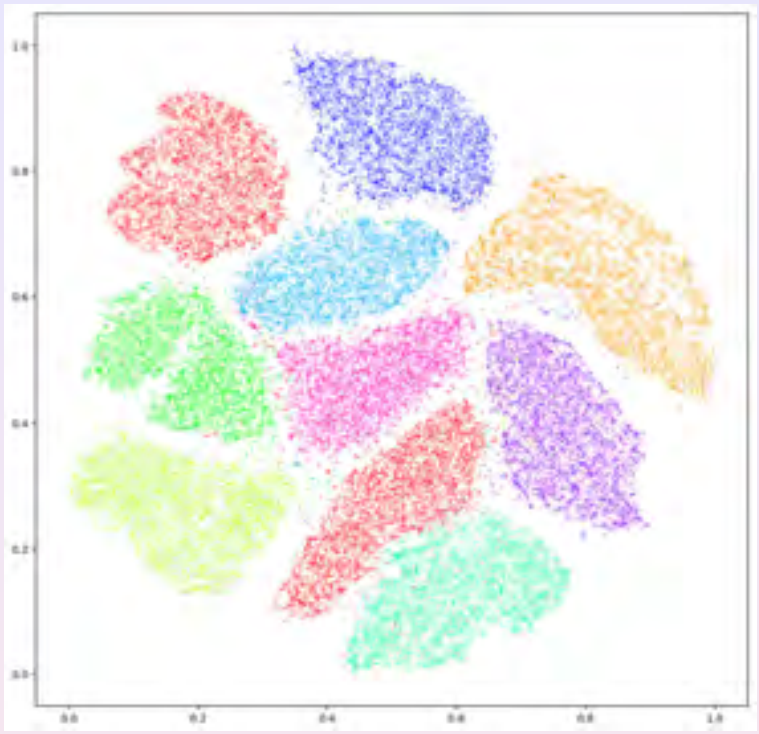
A natural data class can be treated as a probability distribution defined on a low dimensional manifold embedded in a high dimensional ambient space.

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 handwritten digits samples on manifold



b. Hinton's t-SNE embemdding on latent space

- 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.

Hinton's t-SNE Method

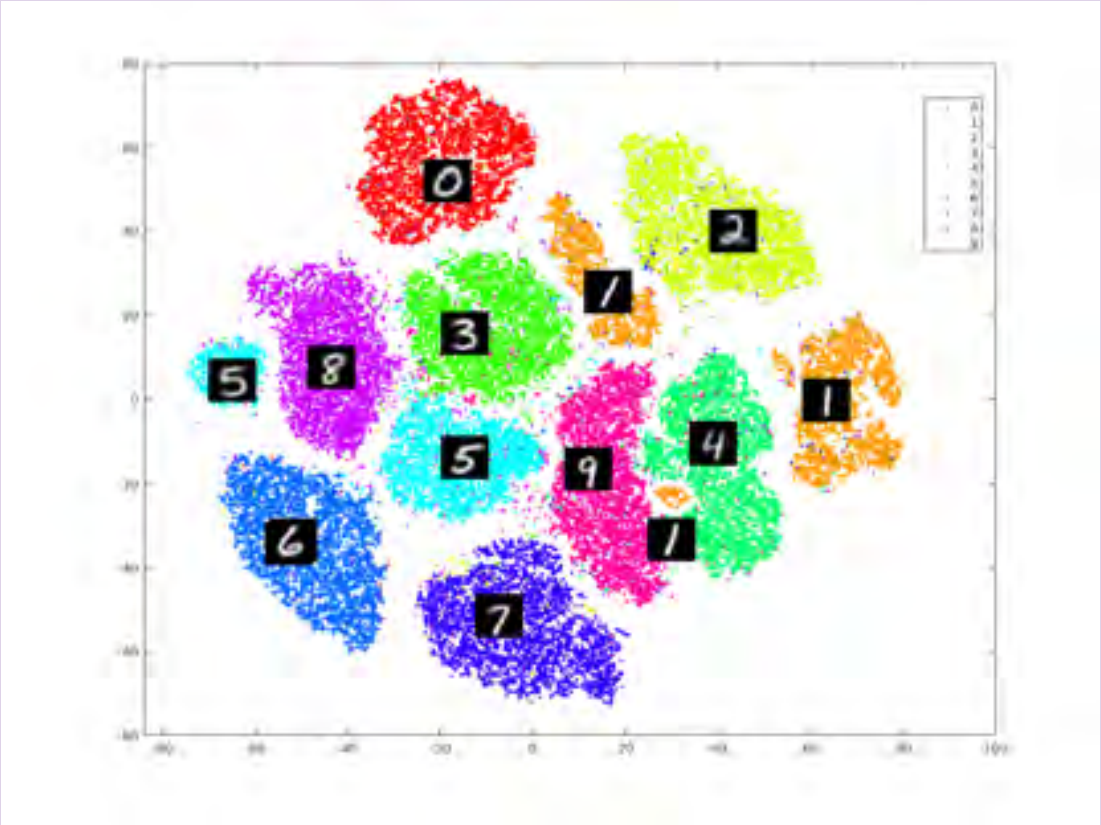
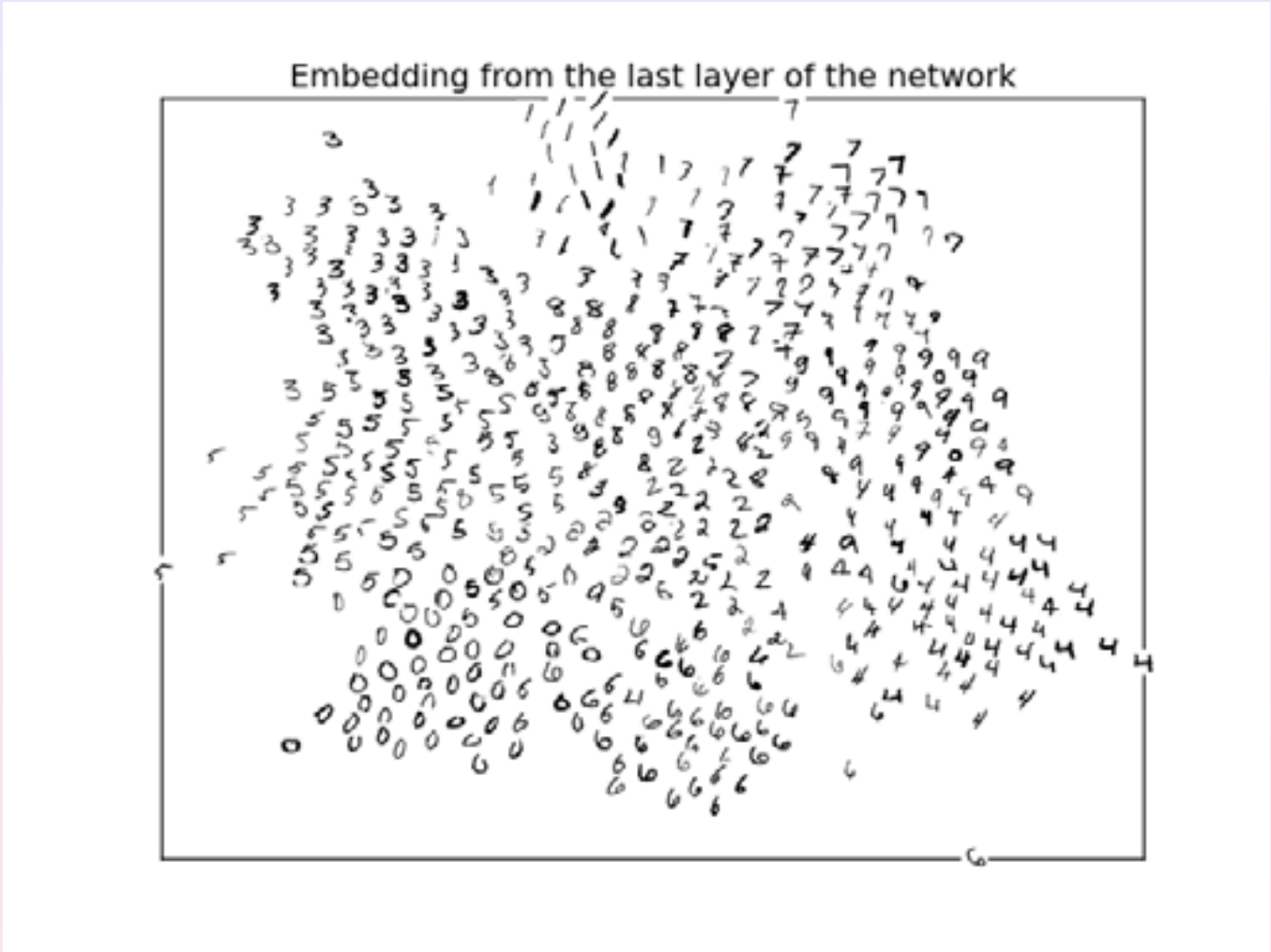


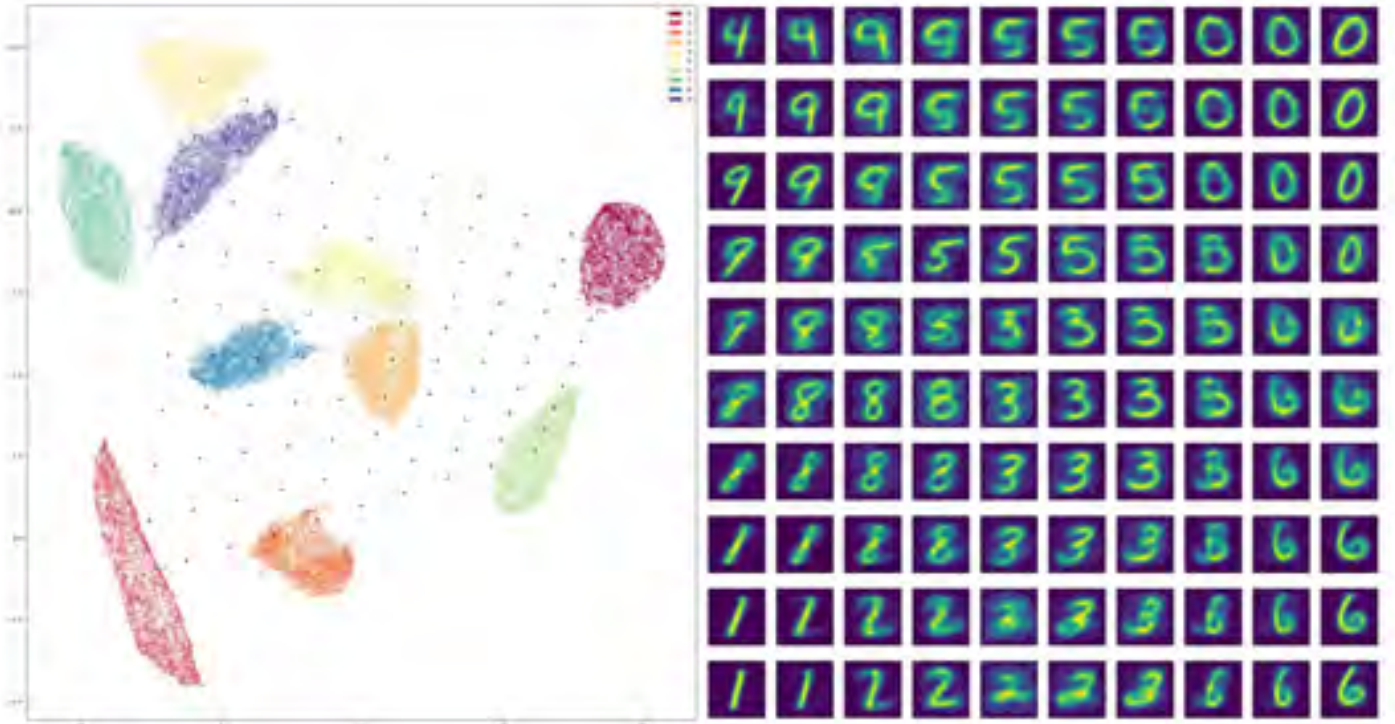
Figure: Each cluster corresponds to a hand written digit. Multiple clusters may correspond to the same digit.

MNIST Siamese Embedding



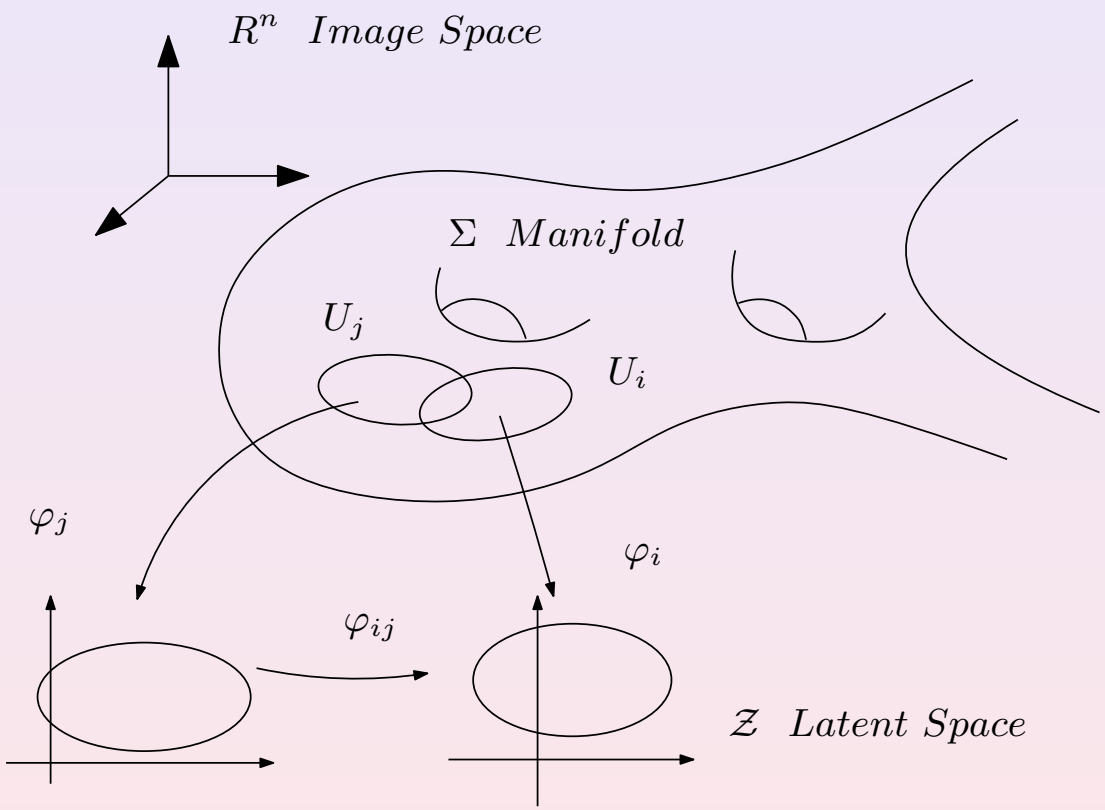
Different embedding result with inferior quality by a Siamese network.

MNIST UMap Embedding



UMap embedding, the samples between modes produce obscure images, which is called mode mixture.

General Model

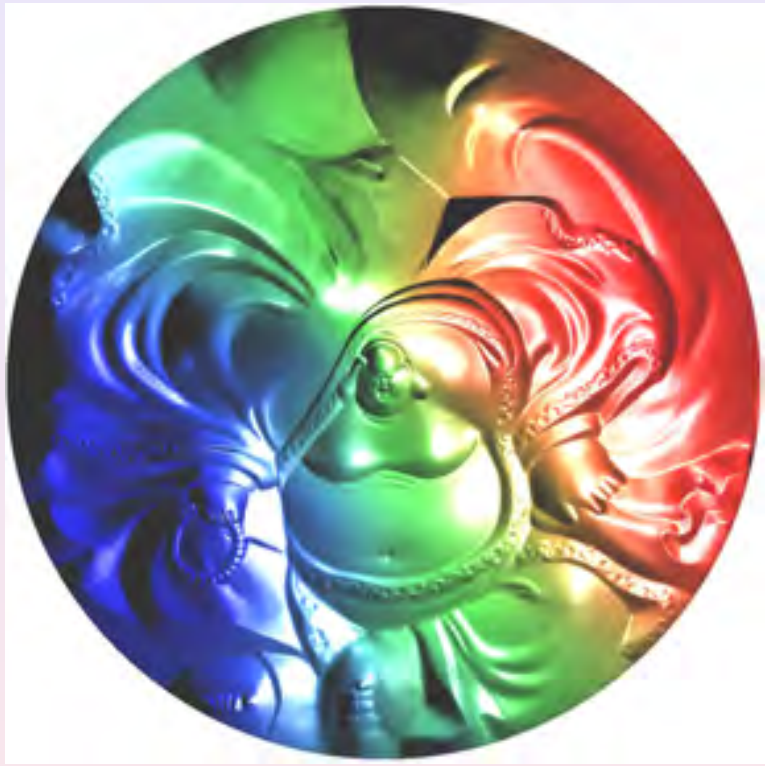


- Ambient Space- image space $\mathbb{R}^n$
- manifold - Support of a distribution μ
- parameter domain - latent space $\mathbb{R}^m$
- coordinates map φ_i- encoding/decoding maps
- φ_{ij} controls the probability measure

Low Dimensional Example

Image space $\mathscr{X}$ is $\mathbb{R}^3$; the data manifold Σ is the happy buddaha.

Example

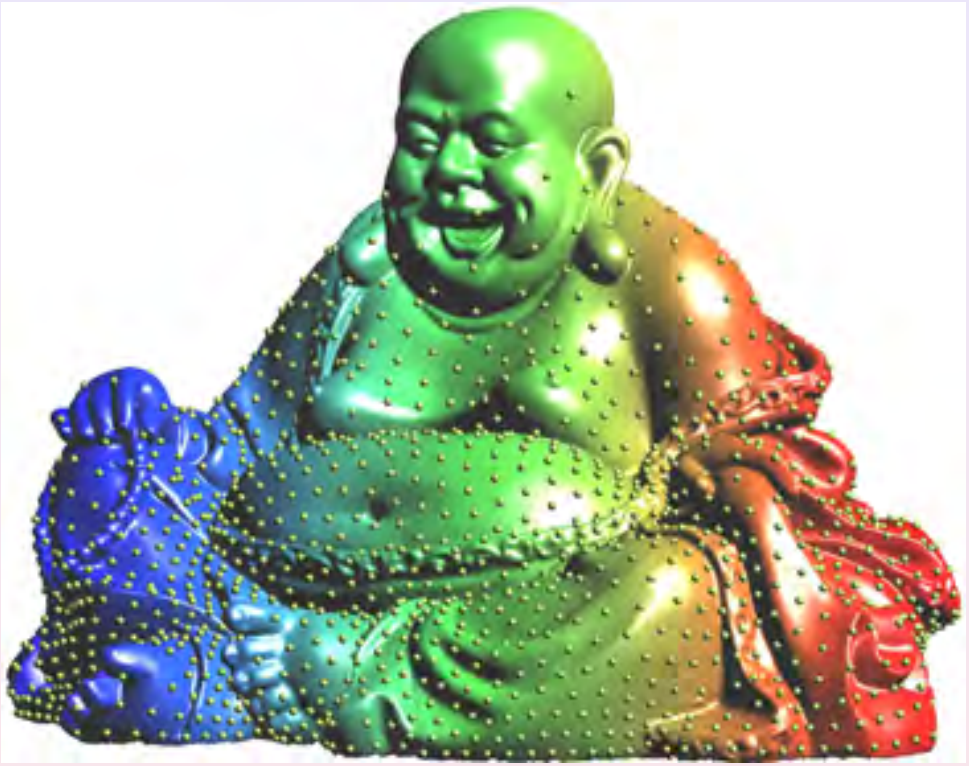


The encoding map is $\varphi_i : \Sigma \to \mathscr{Z}$; the decoding map is $\varphi_i^{-1} : \mathscr{Z} \to \Sigma$.

Example

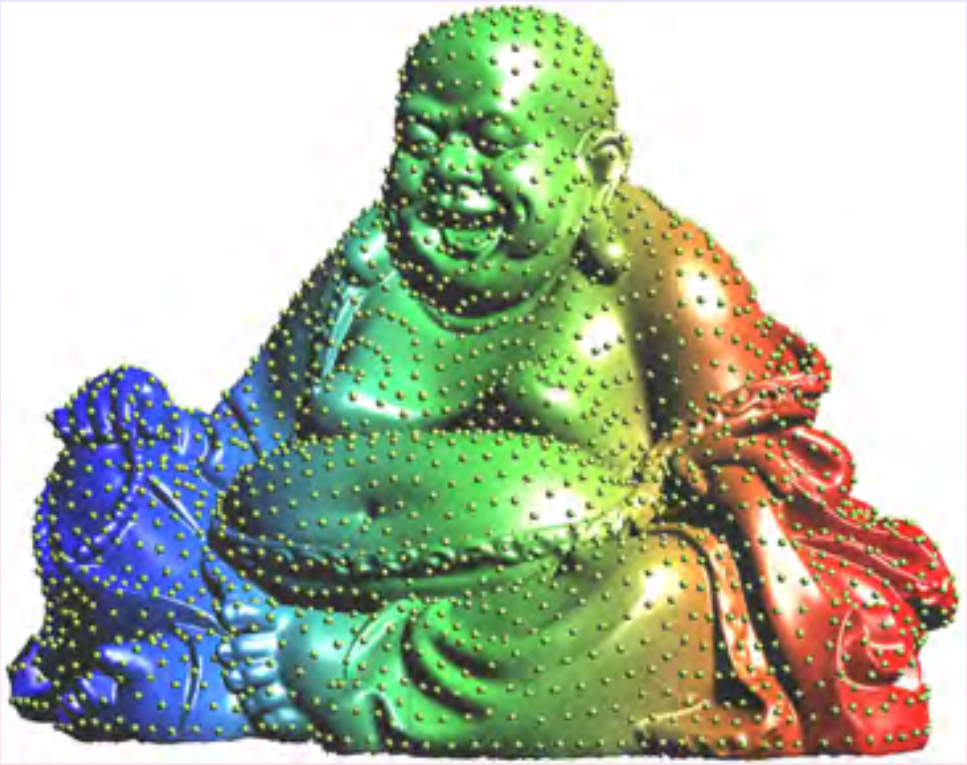
The automorphism of the latent space $\varphi_{ij} : \mathscr{Z} \to \mathscr{Z}$ is the chart transition.

Example



Uniform distribution ζ on the latent space $\mathscr{Z}$, non-uniform distribution on Σ produced by a decoding map.

Example



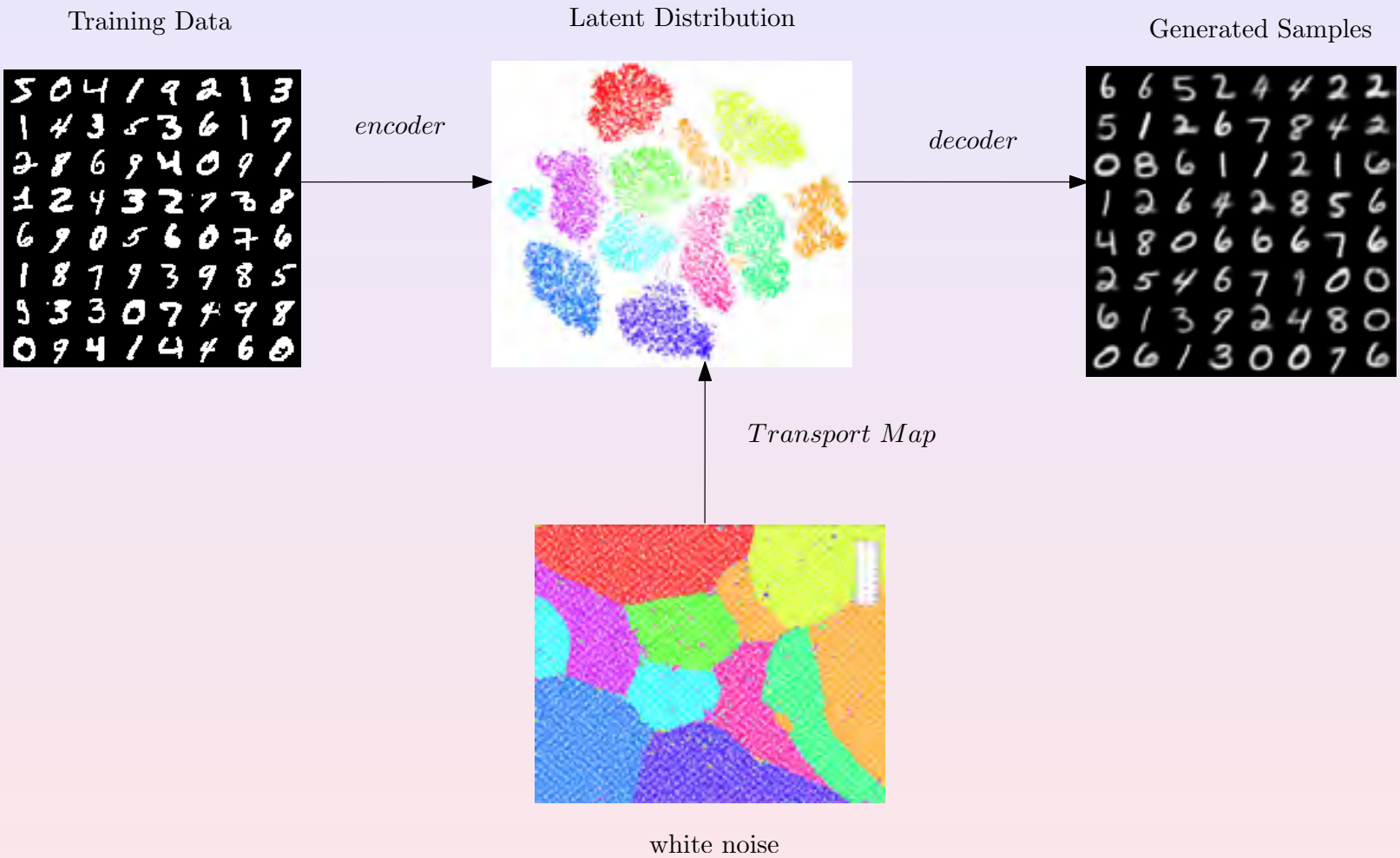
Uniform distribution ζ on the latent space $\mathscr{Z}$, uniform distribution on Σ produced by another decoding map.

Central Tasks for DL

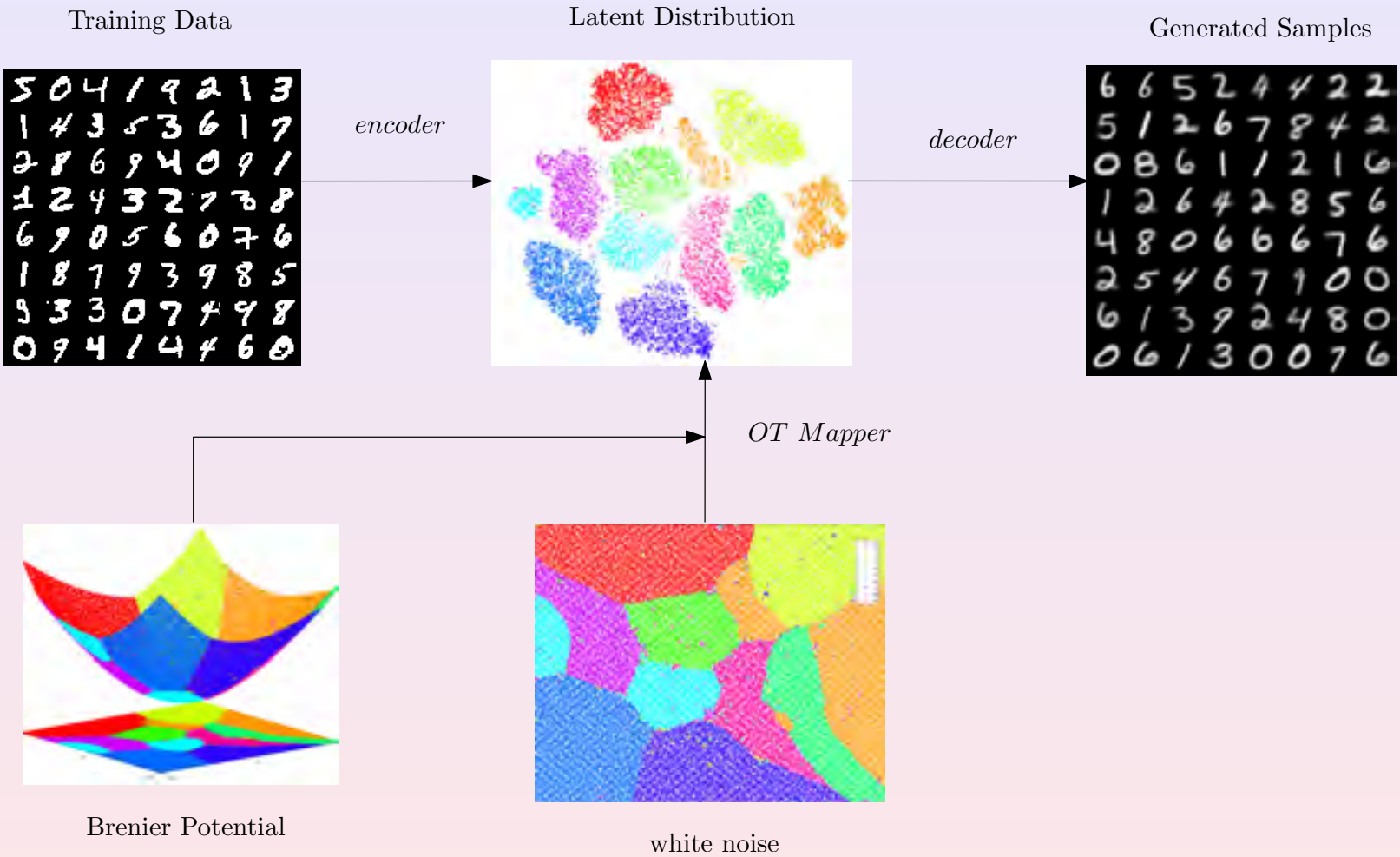
Central Tasks

1. Learn the manifold structure;
2. Learn the probability distribution.

Generative Model Framework



AE-OT Framework



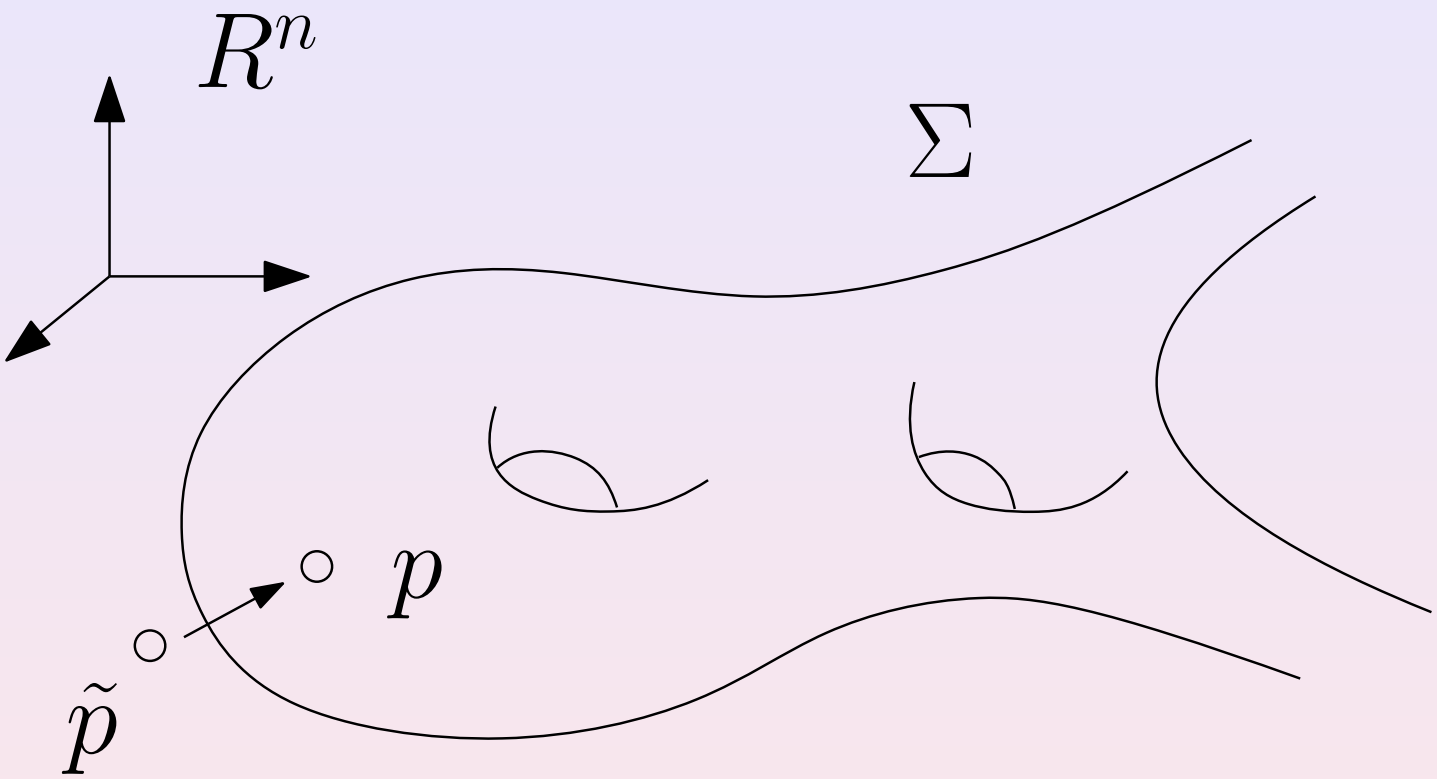
Human Facial Image Manifold

One facial image is determined by a finite number of genes, lighting conditions, camera parameters, therefore all facial images form a manifold.

Manifold view of Generative Model

Given a parametric representation $\varphi : \mathscr{Z} \to \Sigma$, randomly generate a parameter $z \in \mathscr{Z}$ (white noise), $\varphi(z) \in \Sigma$ is a human facial image.

Manifold view of Denoising



Suppose $\tilde{p}$ is a point close to the manifold, $p \in \Sigma$ is the closest point of $\tilde{p}$. The projection $\tilde{p} \to p$ can be treated as denoising.

Manifold view of Denoising

Σ is the clean facial image manifold; noisy image $\tilde{p}$ is a point close to Σ; the closest point $p \in \Sigma$ is the resulting denoised image.

Manifold view of Denoising

Traditional Method

Fourier transform the noisy image, filter out the high frequency component, inverse Fourier transform back to the denoised image.

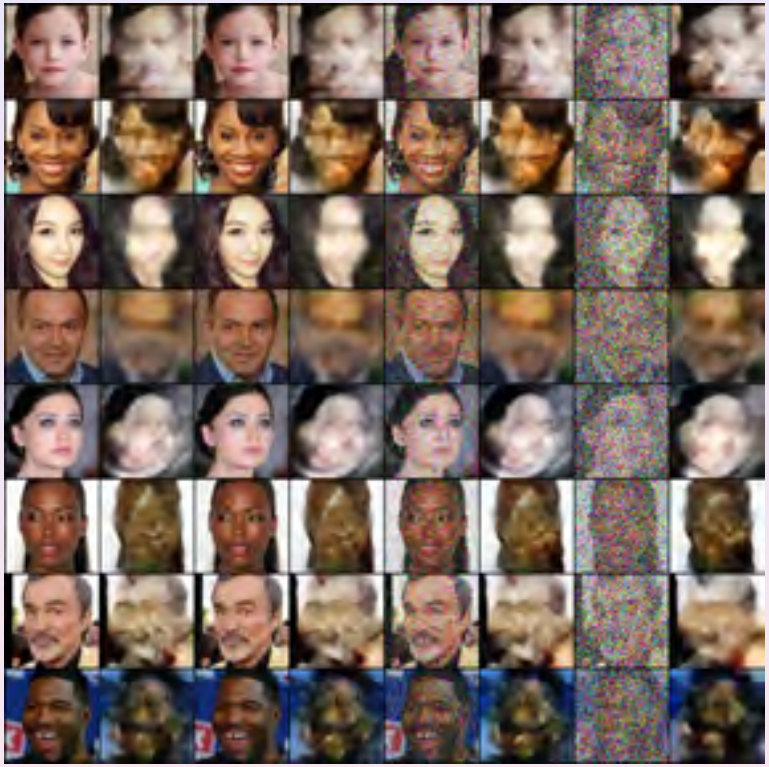
ML Method

Use the clean facial images to train the neural network, obtain a representation of the manifold. Project the noisy image to the manifold, the projection point is the denoised image.

Key Difference

Traditional method is independent of the content of the image; ML method heavily depends on the content of the image. The prior knowledge is encoded by the manifold.

Manifold view of Denoising



If the wrong manifold is chosen, the denoising result is of non-sense. Here we use the cat face manifold to denoise a human face image, the result looks like a cat face.

Manifold Learning

Lemma (Urysohn's Lemma)

Let A,B be closed subsets of a normal topological space X. There exists a continuous function $f : X \to [0,1]$ *such that* $f(A) = 0$ *and* $f(B) = 1$.

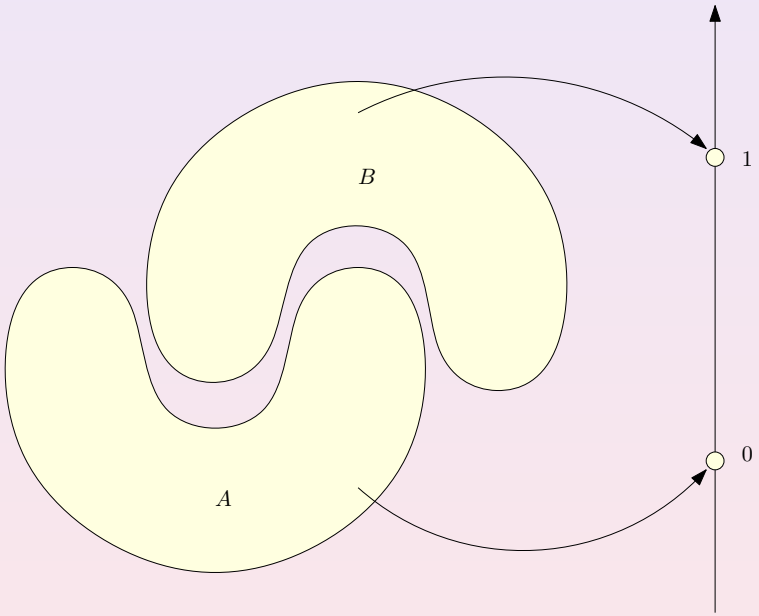


Figure: Urysohn's lemma provides the theoretic tool for pattern recognition, supervised learning.

Theorem (General Position Theorem)

Any m-manifold unknots in $\mathbb{R}^n$ provided $n \geq 2m+2$.

In deep learning, this means that data is easier to manipulate once it is embedded in higher dimensions.

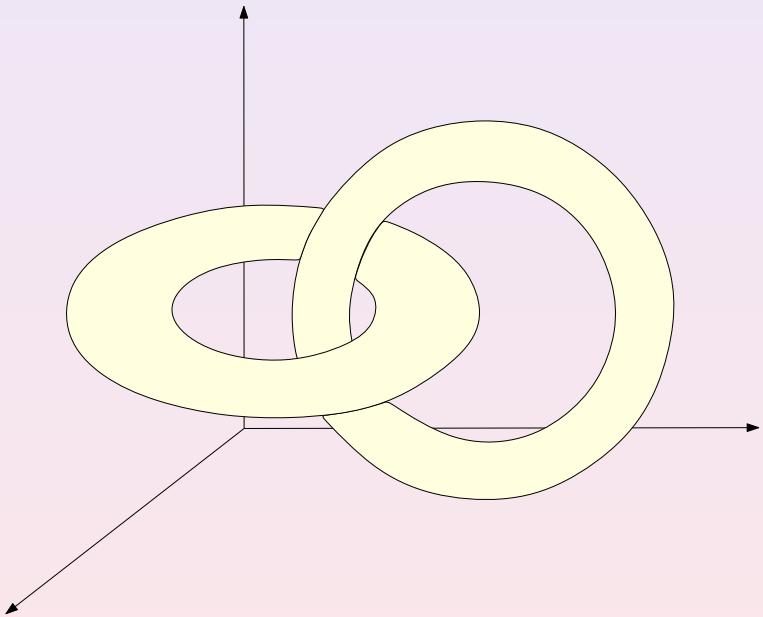


Figure: Increase the dimension of the embedding space to unlink the manifolds.

Whitney Manifold Embedding

Theorem (Whiteny Embedding)

Any smooth real m-dimensional manifold (required also to be Hausdorff and second-countable) can be smoothly embedded in the real 2m-space ($\mathbb{R}^{2m}$).

1. Construct an atlas, $M \subset \bigcup_{i=1}^{k} U_i$;
2. Build a partition of unity;
3. Embed each open set in $\mathbb{R}^m$;
4. Glue the local embeddings to embed M in $\mathbb{R}^{km}$;
5. Random projection to lower dimension.

Universal Approximation

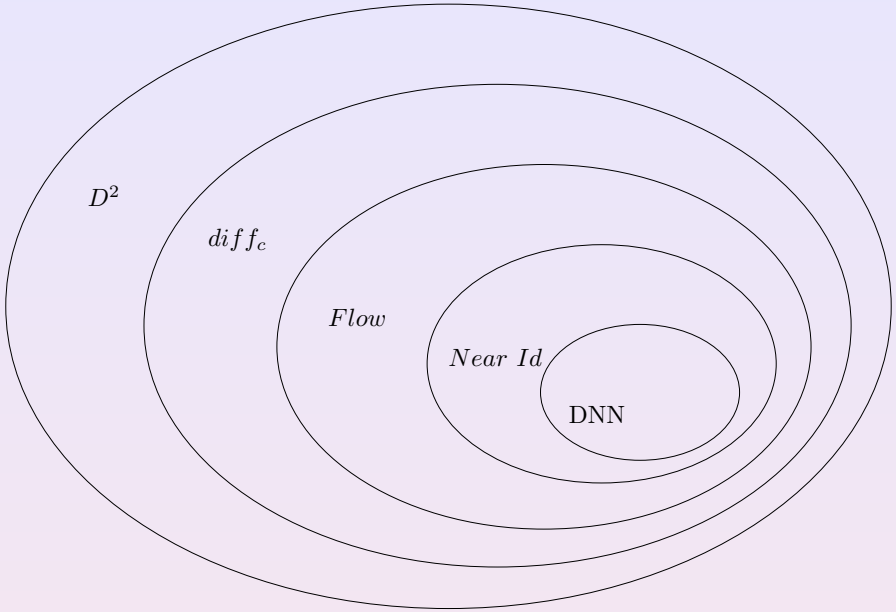
Related to the Hilbert 13th Problem,

Theorem (Kolmogorov-Arnold Representation)

f is a multivariate continuous function, then f can be written as a finite composition of continuous functions of a single variable and the binary operation of addition.

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \varphi_{p,q}(x_p) \right)$$

Universal Approximation



Construct a sequence of nested mapping spaces, $\mathscr{F}_{k+1}$ is simpler than $\mathscr{F}_k$,

$$\mathscr{F}_0 \supset \mathscr{F}_1 \supset \mathscr{F}_2 \cdots \supset \mathscr{F}_n,$$

each mapping $f \in \mathscr{F}_k$ can be approximated by a finite composition of mappings $g_1, g_2, \ldots, g_r \in \mathscr{F}_{k+1}$, $f = g_1 \circ g_2 \circ g_3 \cdots g_r$. $\mathscr{F}_n$ can be computed by deep neural networks.

Universal Approximation

General C^2 diffeomorphisms can be approximated by the followings steps:

1. By finite composition of diffeomorphisms with compact support;
2. By finite composition of flow mappings φ_s:

$$\frac{d}{dt}\varphi(p,t) = \mathbf{v}(\varphi(p,t),t), \quad \varphi(p,0) = id;$$

3. By finite composition of near id maps

$$|Dg - I| < \delta;$$

4. By deep neural networks, such as affine coupling flows etc, which grantee the mapping is invertible;

Autoencoder

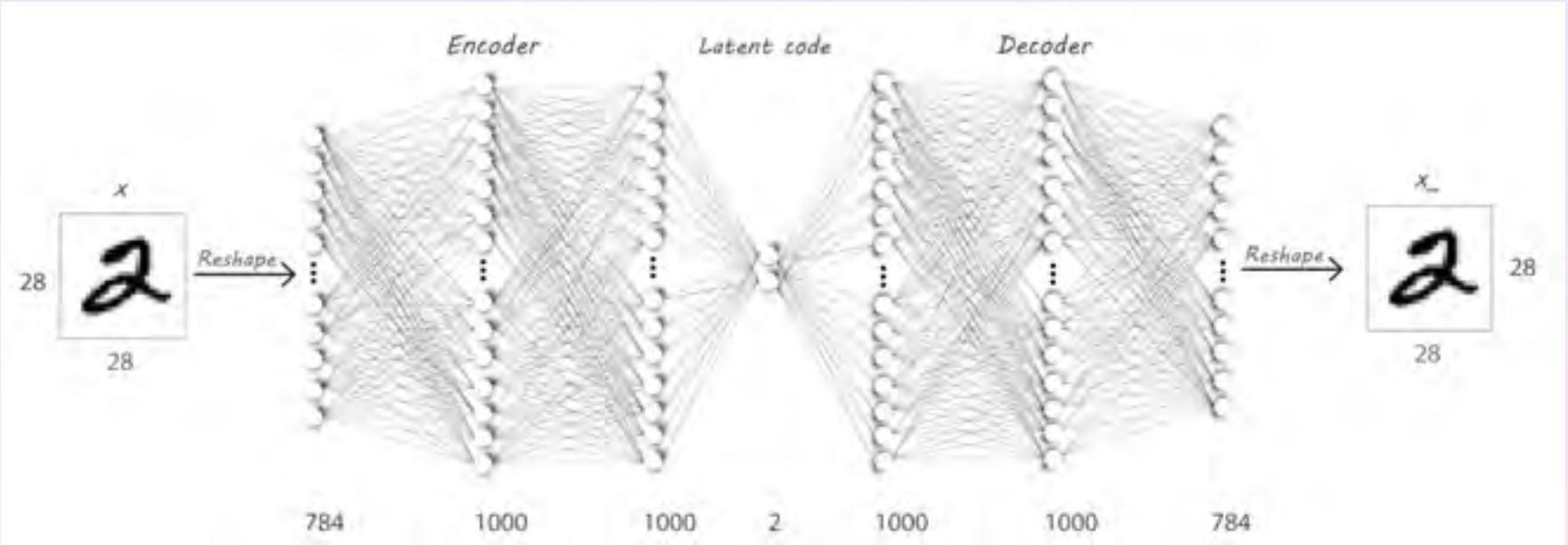


Figure: Auto-encoder architecture.

Ambient space $\mathscr{X}$, latent space $\mathscr{Z}$, encoding map $\varphi_\theta : \mathscr{X} \to \mathscr{Z}$, decoding map $\psi_\theta : \mathscr{Z} \to \mathscr{X}$.

Encoding/Decoding

a. Input manifold $M \subset \mathscr{X}$

b. latent representation $D = \varphi_\theta(M) \subset \mathscr{Z}$

c. reconstructed mfld $\tilde{M} = \psi_\theta(D) \subset \mathscr{X}$

Figure: Auto-encoder pipeline.

Definition (ReLU DNN)

For any number of hidden layers $k \in \mathbb{N}$, input and output dimensions $w_0, w_{k+1} \in \mathbb{N}$, a $\mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ ReLU DNN is given by specifying a sequence of k natural numbers $w_1, w_2, \ldots, w_k$ representing widths of the hidden layers, a set of k affine transformations $T_i : \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$ for $i = 1, \ldots, k$ and a linear transformation $T_{k+1} : \mathbb{R}^{w_k} \to \mathbb{R}^{w_{k+1}}$ corresponding to weights of hidden layers.

The mapping $\varphi_\theta : \mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ represented by this ReLU DNN is

$$\varphi = T_{k+1} \circ \sigma \circ T_k \circ \cdots \circ T_2 \circ \sigma \circ T_1, \tag{1}$$

where $\circ$ denotes mapping composition, θ represent all the weight and bias parameters.

Activated Path

Fix the encoding map φ_θ , let the set of all neurons in the network is denoted as $\mathscr{S}$, all the subsets is denoted as $2^{\mathscr{S}}$.

Definition (Activated Path)

Given a point $\mathbf{x} \in \mathscr{X}$, the *activated path* of $\mathbf{x}$ consists all the activated neurons when $\varphi_\theta(\mathbf{x})$ is evaluated, and denoted as $\rho(\mathbf{x})$. Then the activated path defines a set-valued function $\rho : \mathscr{X} \to 2^{\mathscr{S}}$.

Cell Decomposition

Definition (Cell Decomposition)

Fix a encoding map φ_θ represented by a ReLU DNN, two data points $\mathbf{x}_1, \mathbf{x}_2 \in \mathscr{X}$ are *equivalent*, denoted as $\mathbf{x}_1 \sim \mathbf{x}_2$, if they share the same activated path, $\rho(\mathbf{x}_1) = \rho(\mathbf{x}_2)$. Then each equivalence relation partitions the ambient space $\mathscr{X}$ into cells,

$$\mathscr{D}(\varphi_\theta) : \mathscr{X} = \bigcup_\alpha U_\alpha,$$

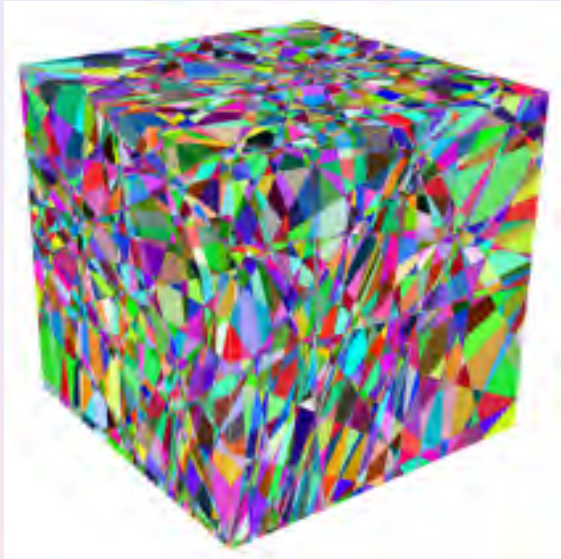
each equivalence class corresponds to a cell: $\mathbf{x}_1, \mathbf{x}_2 \in U_\alpha$ if and only if $\mathbf{x}_1 \sim \mathbf{x}_2$. $\mathscr{D}(\varphi_\theta)$ is called the cell decomposition induced by the encoding map φ_θ.

Furthermore, φ_θ maps the cell decomposition in the ambient space $\mathscr{D}(\varphi_\theta)$ to a cell decomposition in the latent space.

Piecewise Linear Mapping

d. cell decomposition $\mathscr{D}(\varphi_\theta)$

e. latent space cell decomposition



f. cell decomposition $\mathscr{D}(\psi_\theta \circ \varphi_\theta)$

Piecewise linear encoding/decoding maps induce cell decompositions of the ambient space and the latent space.

Learning Capability

Definition (Learning Capability)

Given a ReLU DNN $N(w_0, \ldots, w_{k+1})$, its rectified linear complexity is the upper bound of the number of pieces of all PL functions φ_θ represented by N,

$$\mathcal{N}(N) := \max_\theta \mathcal{N}(\varphi_\theta),$$

where $\mathcal{N}(\varphi_\theta)$ is the number of pieces of the PL function φ_θ.

This gives a measurement for the representation capability of a neural network.

RL Complexity Estimate

Lemma

The maximum number of parts one can get when cutting d-dimensional space $\mathbb{R}^d$ with n hyperplanes is denoted as $\mathscr{C}(d,n)$, then

$$\mathscr{C}(d,n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{d}. \tag{2}$$

Proof.

Suppose n hyperplanes cut $\mathbb{R}^d$ into $\mathscr{C}(d,n)$ cells, each cell is a convex polyhedron. The $(n+1)$-th hyperplane is π, then the first n hyperplanes intersection π and partition π into $\mathscr{C}(d-1,n)$ cells, each cell on π partitions a polyhedron in $\mathbb{R}^d$ into 2 cells, hence we get the formula

$$\mathscr{C}(d,n+1) = \mathscr{C}(d,n) + \mathscr{C}(d-1,n).$$

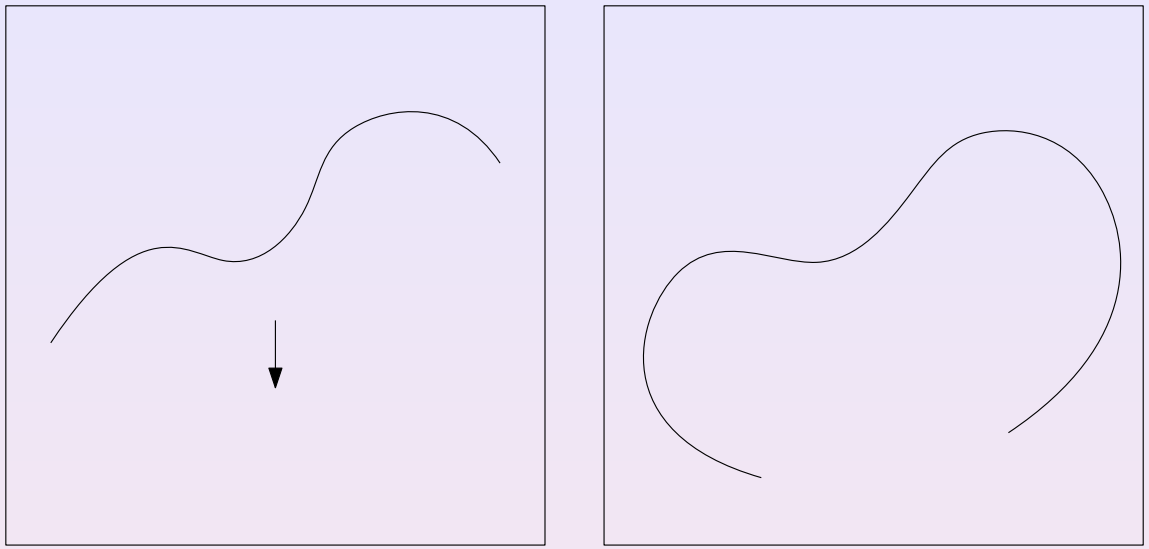
RL Complexity Upper Bound

Theorem (Rectified Linear Complexity of a ReLU DNN)

Given a ReLU DNN $N(w_0, \ldots, w_{k+1})$, representing PL mappings $\varphi_\theta : \mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ with k hidden layers of widths $\{w_i\}_{i=1}^{k}$, then the linear rectified complexity of N has an upper bound,

$$\mathscr{N}(N) \leq \Pi_{i=1}^{k+1} \mathscr{C}(w_{i-1}, w_i). \tag{3}$$

RL Complexity of Manifold



a. linear rectifiable b. non-linear-rectifiable

Definition (Linear Rectifiable Manifold)

Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, we say M is linear rectifiable, if there exists an affine map $\varphi : \mathbb{R}^n \to \mathbb{R}^m$, such that the restriction of φ on M, $\varphi|_M : M \to \varphi(M) \subset \mathbb{R}^m$, is homeomorphic. φ is called the corresponding rectified linear map of M.

Manifold RL Complexity

Definition (Linear Rectifiable Atlas)

Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, $\mathscr{A} = \{(U_\alpha, \varphi_\alpha\}$ is an atlas of M. If each chart $(U_\alpha, \varphi_\alpha)$ is linear rectifiable, $\varphi_\alpha : U_\alpha \to \mathbb{R}^m$ is the rectified linear map of U_α, then the atlas is called a linear rectifiable atlas of M.

Definition (Rectified Linear Complexity of a Manifold)

Suppose M is a m-dimensional manifold embedded in $\mathbb{R}^n$, the rectified linear complexity of M is denoted as $\mathscr{N}(\mathbb{R}^n, M)$ and defined as,

$$\mathscr{N}(\mathbb{R}^n, M) := \min\{|\mathscr{A}| \quad |\mathscr{A} \text{ is a linear rectifiable altas of } M\}. \tag{4}$$

Encodable Condition

Definition (Encoding Map)

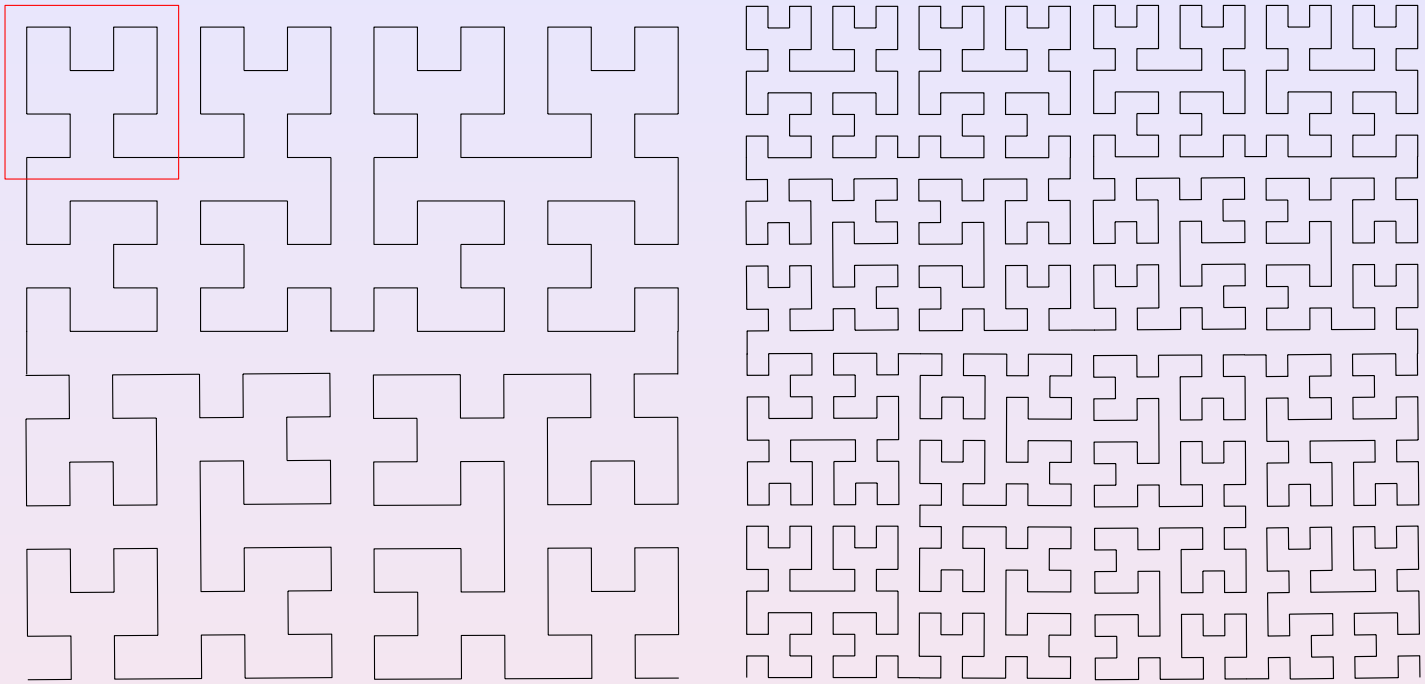
Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, a continuous mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is called an encoding map of $(\mathbb{R}^n, M)$, if restricted on M, $\varphi|_M : M \to \varphi(M) \subset \mathbb{R}^m$ is homeomorphic.

Theorem (Encodable Condition)

Suppose a ReLU DNN $N(w_0, \ldots, w_{k+1})$ represents a PL mapping $\varphi_\theta : \mathbb{R}^n \to \mathbb{R}^m$, M is a m-dimensional manifold embedded in $\mathbb{R}^n$. If φ_θ is an encoding mapping of $(\mathbb{R}^n, M)$, then the rectified linear complexity of N is no less that the rectified linear complexity of $(\mathbb{R}^n, M)$,

$$\mathcal{N}(\mathbb{R}^n, M) \leq \mathcal{N}(\varphi_\theta) \leq \mathcal{N}(N).$$

Representation Limitation Theorem



C_1 Peano curve C_2 Peano curve

Figure: $\mathcal{N}(\mathbb{R}^2, C_n) \geq 4^{n+1}$

Theorem

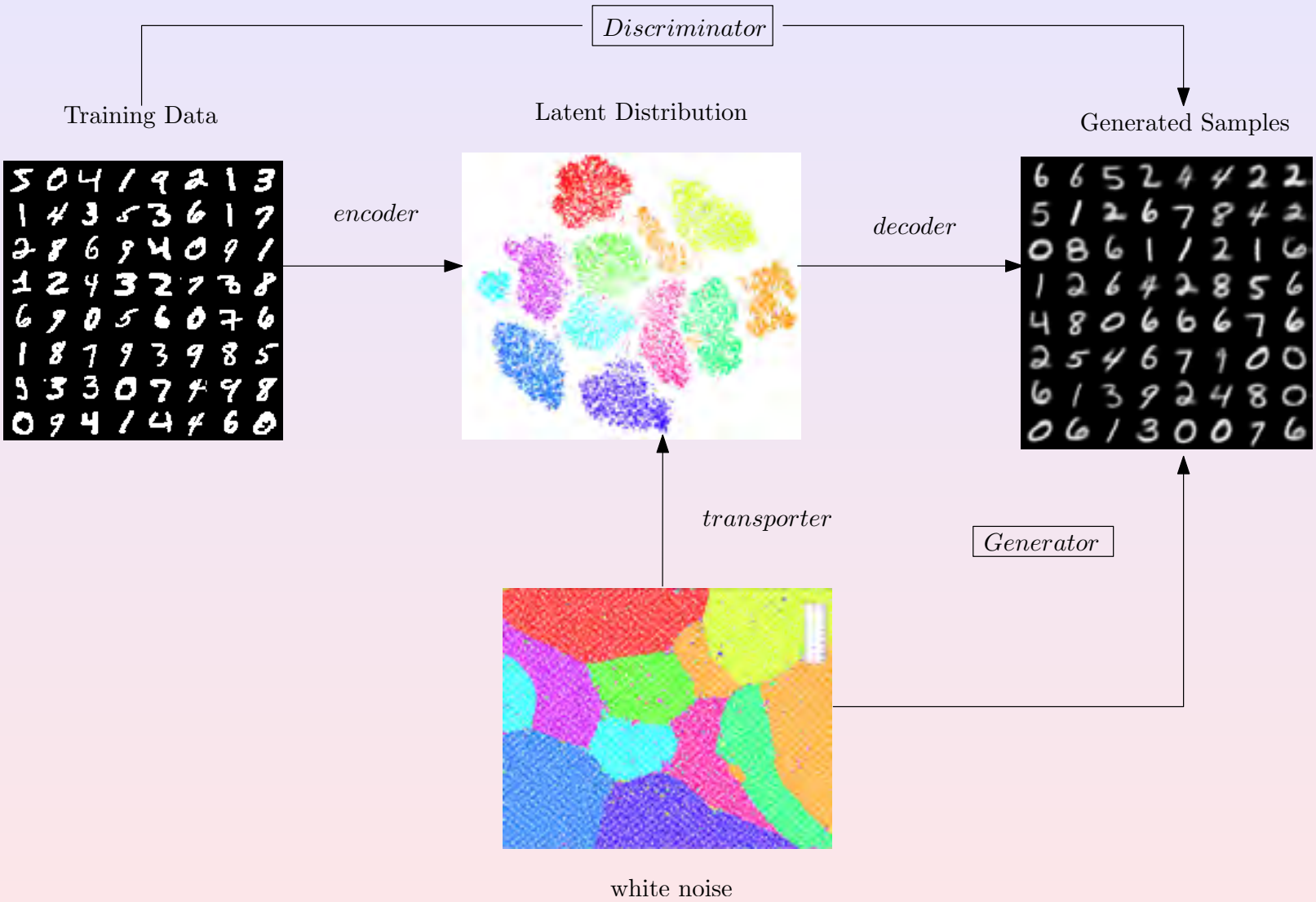
Given any ReLU deep neural network $N(w_0, w_1, \ldots, w_k, w_{k+1})$, there is a manifold M embedded in $\mathbb{R}^{w_0}$, such that M can not be encoded by N.

Probability Measure Learning

Generative Model

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

GAN model


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

GAN Overview

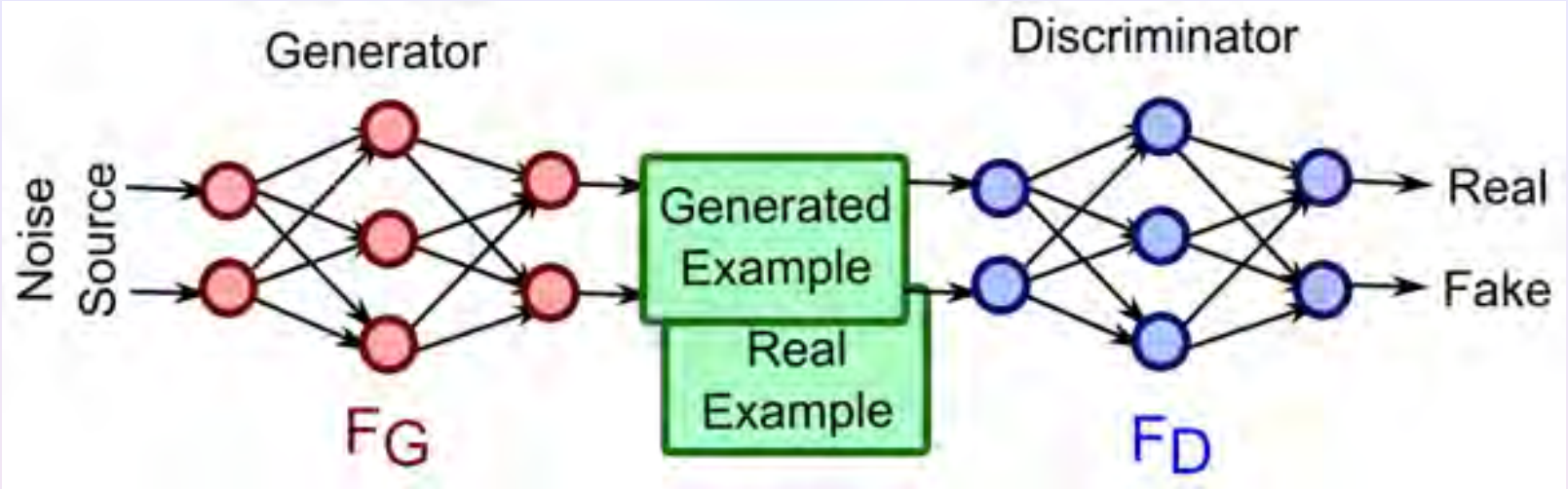


Figure: GAN DNN model.

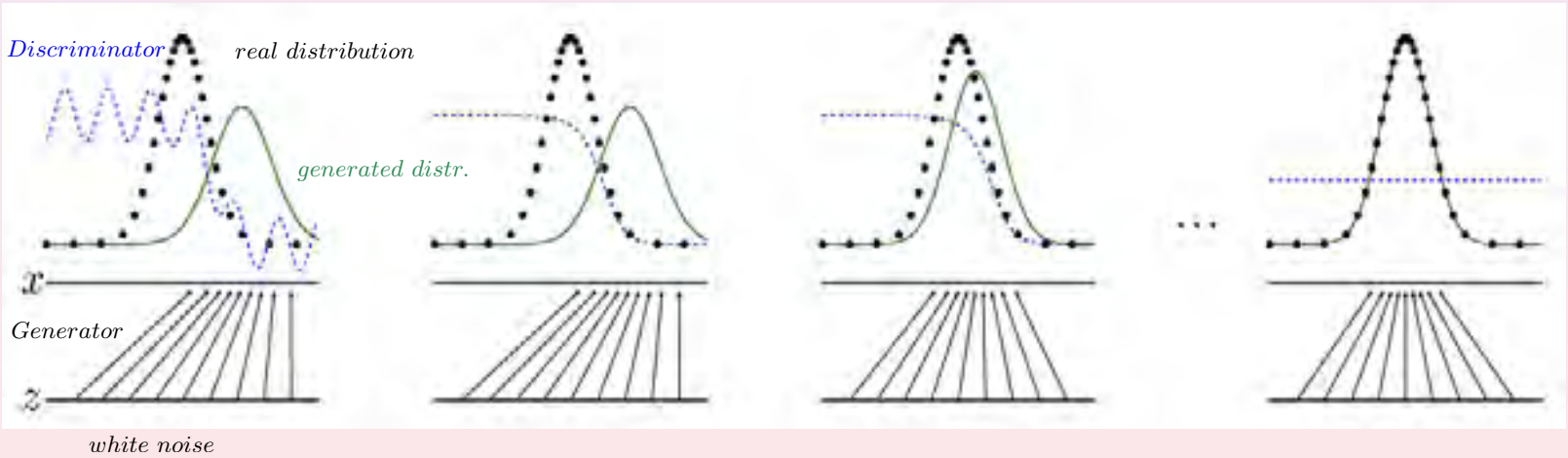
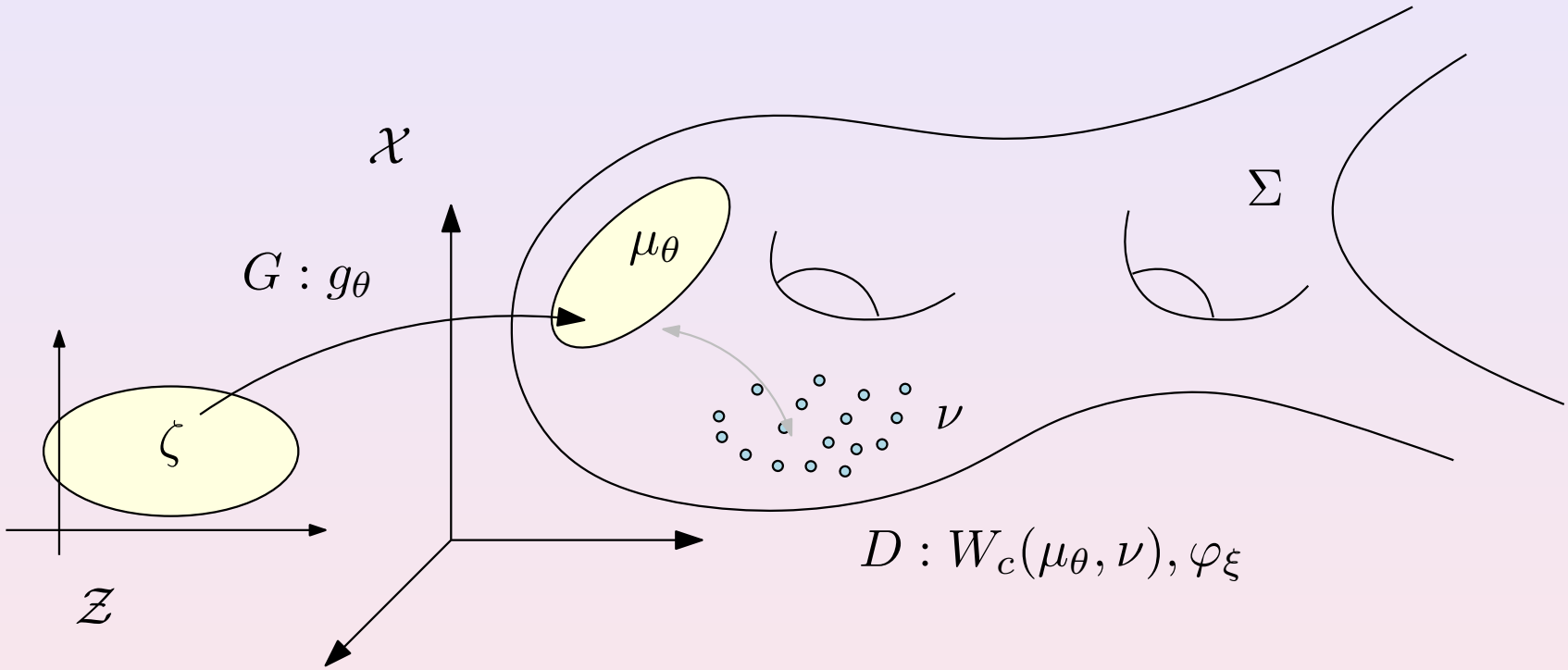


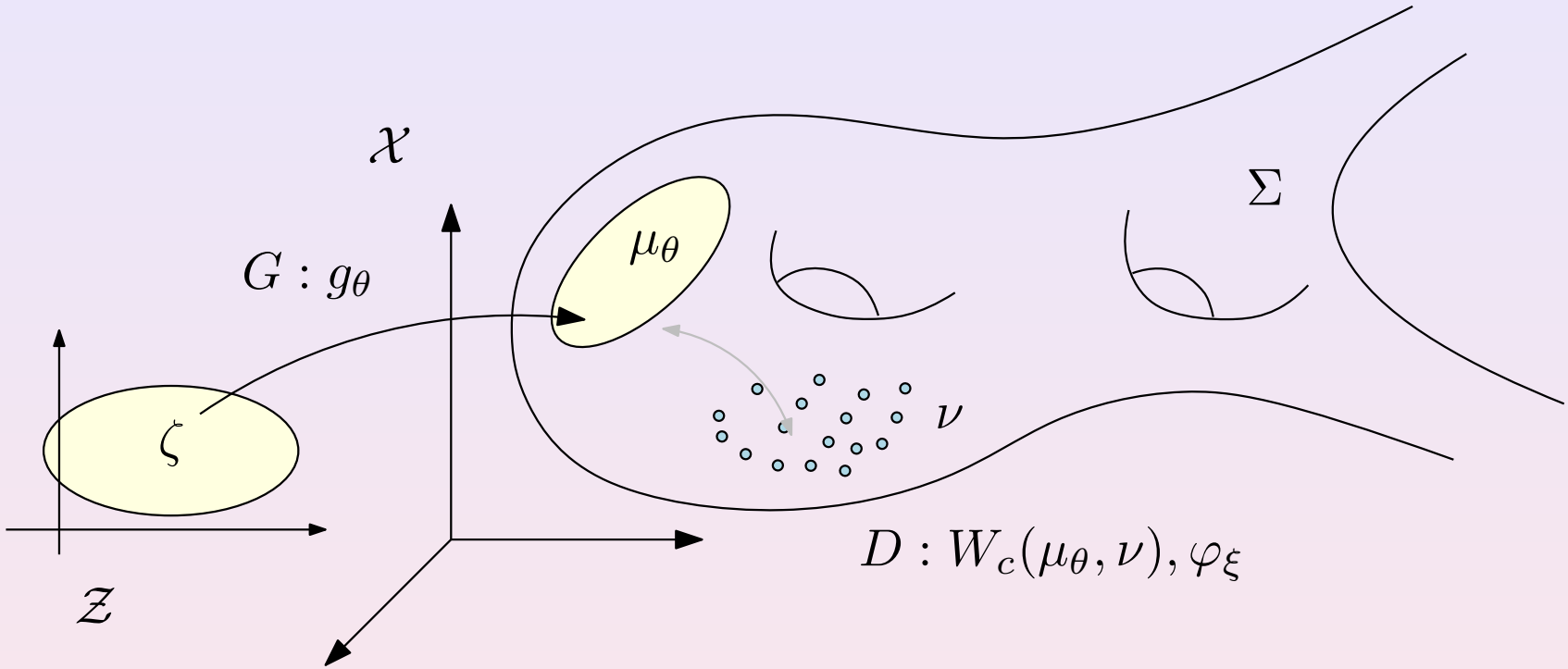
Figure: GAN learning process.

Wasserstein GAN Model



$\mathscr{X}$-image space; Σ-supporting manifold; $\mathscr{Z}$-latent space; $W_c(\cdot,\cdot)$ is the Wasserstein distance.

Wasserstein GAN Model



ν-training data distribution; ζ-uniform distribution;
$\mu_\theta = g_{\theta\#}\zeta$-generated distribution; G - generator computes g_θ;
D -discriminator, measures the Wasserstein distance between ν and μ_θ, $W_c(\mu_\theta, \nu)$.

Optimal Transport Framework

Overview

Motivation

- Given a manifold X, all the probability distributions on X form an infinite dimensional manifold, Wasserstein Space $\mathscr{P}(X)$;
- Deep Learning tasks are reduced to optimization in $\mathscr{P}(X)$, 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 $\mathscr{P}(X)$.

Overview

Solution

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

Overview

- 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

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

Equivalence to Conventional DL Methods

- Entropy function is convex along the geodesics on $\mathscr{P}(X)$;
- The Hessian of entropy defines another Riemannian metric of $\mathscr{P}(X)$;
- 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.

Monge Problem

Problem (Monge)

Find a measure-preserving transportation map $T : (X, \mu) \to (Y, \mu)$ *that minimizes the transportation cost,*

$$(MP) \qquad \min_{T_{\#}\mu=\nu} \mathscr{C}(T) = \min_{T_{\#}\mu=\nu} \int_X c(x, T(x)) d\mu(x).$$

such kind of map is called the optimal mass transportation map.

Definition (Wasserstein distance)

The transportation cost of the optimal transportation map $T : (X, \mu) \to (Y, \nu)$ is called the Wasserstein distance between μ and ν, denoted as

$$W_c^2(\mu, \nu) := \min_{T_{\#}\mu=\nu} \mathscr{C}(T).$$

Kantorovich Problem

Kantorovich relaxed transportation maps to transportation schemes.

Problem (Kantorovich)

Find an optimal transportation scheme, namely a joint probability measure $\rho \in \mathscr{P}(X \times Y)$, with marginal measures $\rho_{x\#} = \mu$, $\rho_{y\#} = \nu$, that minimizes the transportation cost,

$$(KP) \qquad \min_{\rho} \left\{ \int_{X \times Y} c(x,y) d\rho(x,y) \,\middle|\, \rho_{x\#} = \mu,\ \rho_{y\#} = \nu \right\}.$$

Kantorovich solved this problem by inventing linear programming, and won Nobel's prize in economics in 1975.

Kantorovich Dual Problem

By the duality of linear programming, Kantorovich problem has the dual form:

Problem (Kantorovich Dual)

Find an functions $\varphi : X \to \mathbb{R}$ *and* $\psi : Y \to \mathbb{R}$*, such that*

$$(DP) \quad \max_{\varphi,\psi} \left\{ \int_X \varphi(x) du(x) + \int_Y \psi(y) dv(y), \varphi(x) + \psi(y) \leq c(x,y) \right\}.$$

Kantorovich Dual Problem

Definition (c-transformation)

Given a function $\varphi : X \to \mathbb{R}$, and $c(x, y) : X \times Y \to \mathbb{R}$, its c-transform $\varphi^c : Y \to \mathbb{R}$ is given by

$$\varphi^c(y) := \inf_{x \in X} \{c(x, y) - \varphi(x)\}.$$

Problem (Kantorovich Dual)

The Kantorovich Dual problem can be reformulated as

$$(DP) \qquad \max_{\varphi} \left\{ \int_X \varphi(x) du(x) + \int_Y \varphi^c(y) dv(y) \right\}.$$

φ is called Kantorovich potential.

Brenier's Approach

Theorem (Brenier)

If $\mu, \nu > 0$ and X is convex, and the cost function is quadratic distance,

$$c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}|\mathbf{x} - \mathbf{y}|^2$$

then there exists a convex function $u : X \to \mathbb{R}$ unique upto a constant, such that the unique optimal transportation map is given by the gradient map

$$T : \mathbf{x} \to \nabla u(\mathbf{x}).$$

Problem (Brenier)

Find a convex function $u : X \to \mathbb{R}$, such that

$$(BP) \qquad (\nabla u)_{\#}\mu = \nu,$$

u is called the Brenier potential.

Brenier's Approach

From Jacobian equation, one can get the necessary condition for Brenier potential.

Problem (Brenier)

Find the C^2 Brenier potential $u : X \to \mathbb{R}$ statisfies the Monge-Ampere equation

$$(BP) \qquad det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right) = \frac{\mu(\mathbf{x})}{\nu(\nabla f(\mathbf{x}))}.$$

Convex Geometry

Minkowski problem - General Case

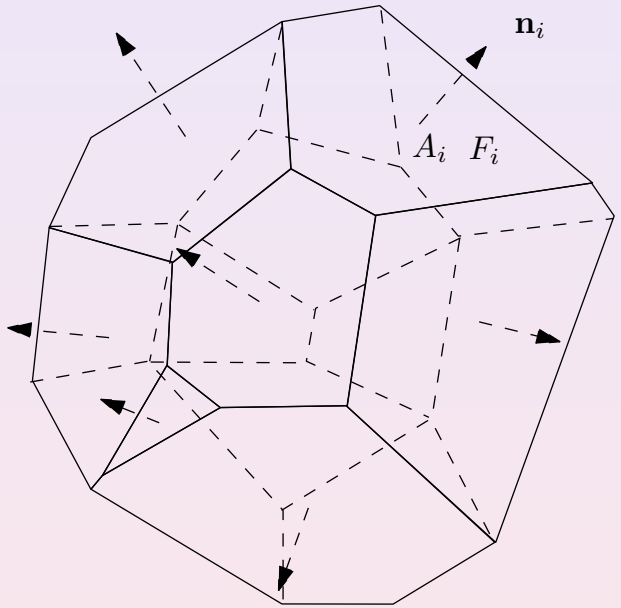
Minkowski Problem

Given k unit vectors $\mathbf{n}_1, \cdots, \mathbf{n}_k$ not contained in a half-space in $\mathbb{R}^n$ and $A_1, \cdots, A_k > 0$, such that

$$\sum_i A_i \mathbf{n}_i = \mathbf{0},$$

find a compact convex polytope P with exactly k codimension-1 faces $F_1, \cdots, F_k$, such that

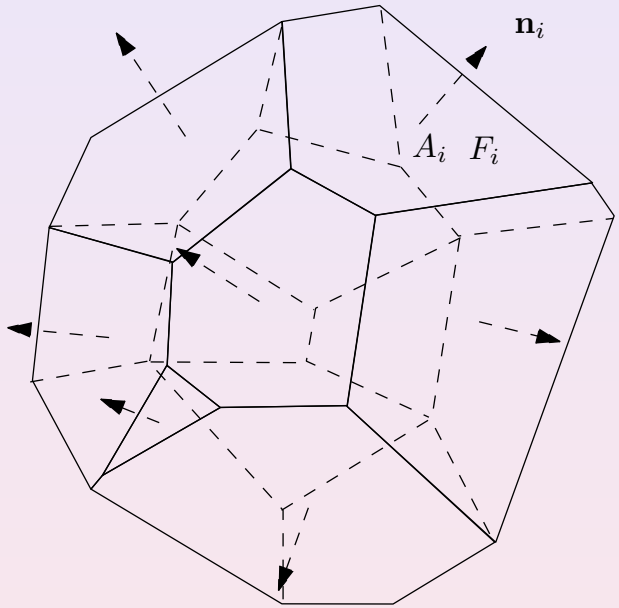
1. $area(F_i) = A_i$,
2. $\mathbf{n}_i \perp F_i$.



Minkowski problem - General Case

Theorem (Minkowski)

P exists and is unique up to translations.



Alexandrov Theorem

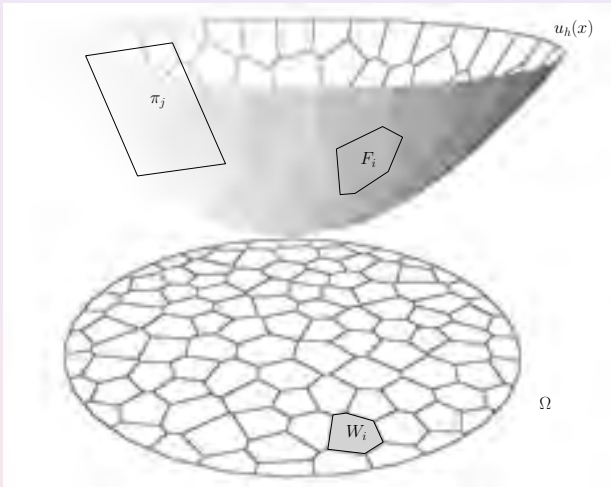
Theorem (Alexandrov 1950)

Given Ω compact convex domain in $\mathbb{R}^n$, $p_1, \cdots, p_k$ distinct in $\mathbb{R}^n$, $A_1, \cdots, A_k > 0$, such that $\sum A_i = Vol(\Omega)$, there exists PL convex function

$$f(\mathbf{x}) := \max\{\langle \mathbf{x}, \mathbf{p}_i \rangle + h_i | i = 1, \cdots, k\}$$

unique up to translation such that

$$Vol(W_i) = Vol(\{\mathbf{x} | \nabla f(\mathbf{x}) = \mathbf{p}_i\}) = A_i.$$



Alexandrov's proof is topological, not variational. It has been open for years to find a constructive proof.

Variational Proof

Theorem (Gu-Luo-Sun-Yau 2013)

Ω is a compact convex domain in $\mathbb{R}^n$, $y_1, \cdots, y_k$ distinct in $\mathbb{R}^n$, μ a positive continuous measure on Ω. For any $\nu_1, \cdots, \nu_k > 0$ with $\sum \nu_i = \mu(\Omega)$, there exists a vector $(h_1, \cdots, h_k)$ so that

$$u(\mathbf{x}) = \max\{\langle \mathbf{x}, \mathbf{p}_i \rangle + h_i\}$$

satisfies $\mu(W_i \cap \Omega) = \nu_i$, where $W_i = \{\mathbf{x} | \nabla f(\mathbf{x}) = \mathbf{p}_i\}$. Furthermore, $\mathbf{h}$ is the maximum point of the convex function

$$E(\mathbf{h}) = \sum_{i=1}^{k} \nu_i h_i - \int_{\mathbf{0}}^{\mathbf{h}} \sum_{i=1}^{k} w_i(\eta) d\eta_i,$$

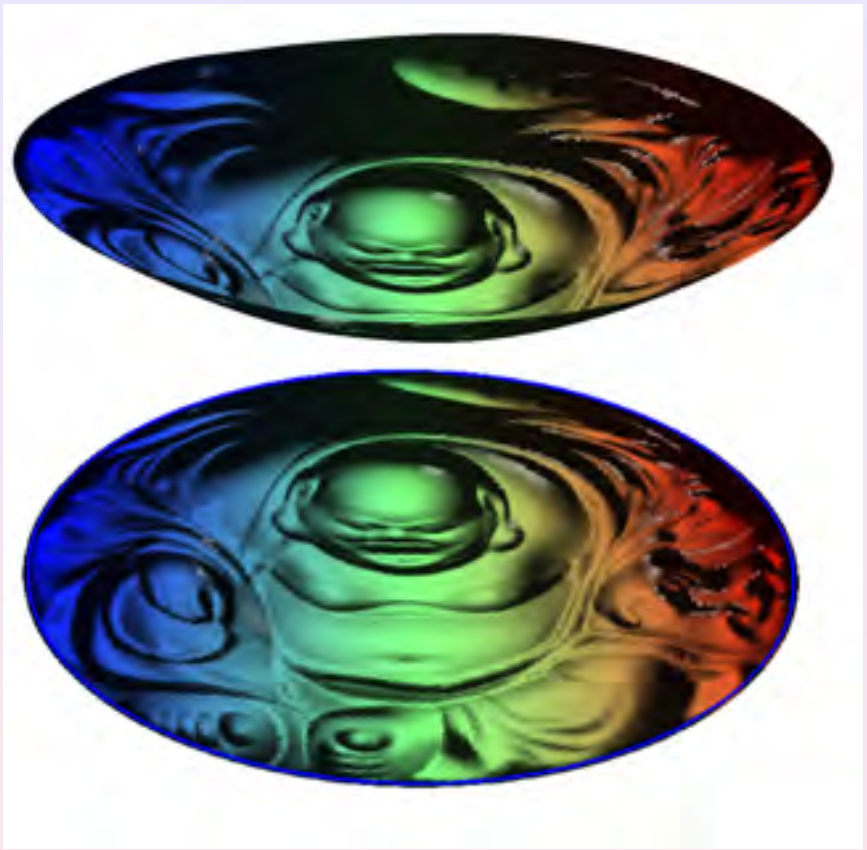
where $w_i(\eta) = \mu(W_i(\eta) \cap \Omega)$ is the μ-volume of the cell.

Variational Proof

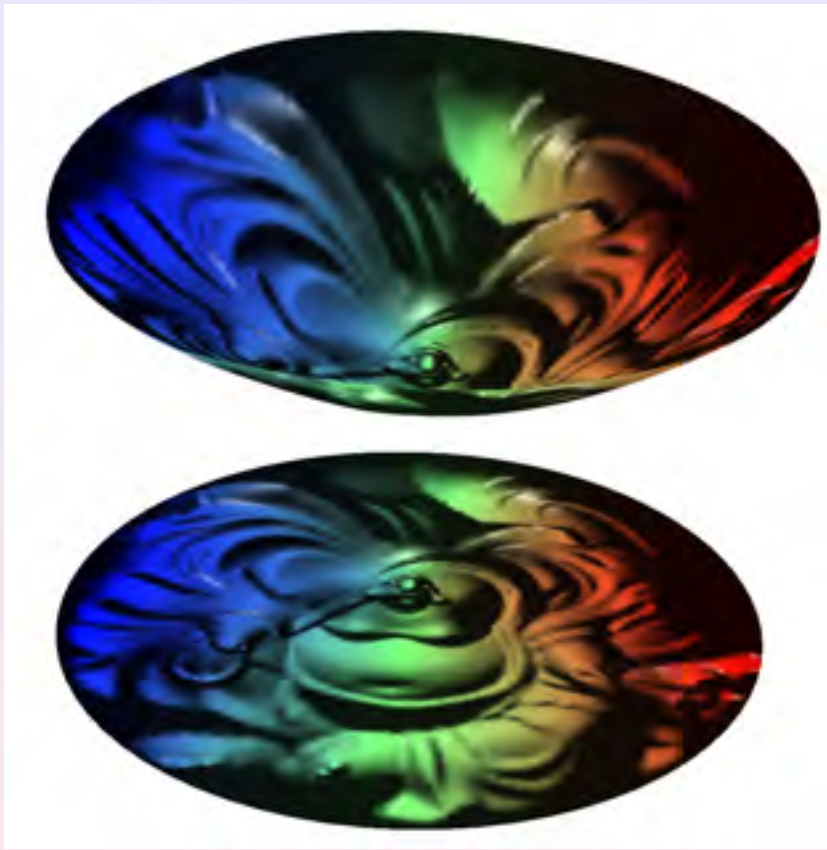
X. Gu, F. Luo, J. Sun and S.-T. Yau, “Variational Principles for Minkowski Type Problems, Discrete Optimal Transport, and Discrete Monge-Ampere Equations”, arXiv:1302.5472

Accepted by Asian Journal of Mathematics (AJM)

Computational Algorithm



a. Brenier potential



b. Legendre dual

Figure: Brenier potential and its Legendre dual for the Buddha example.

Computational Algorithm

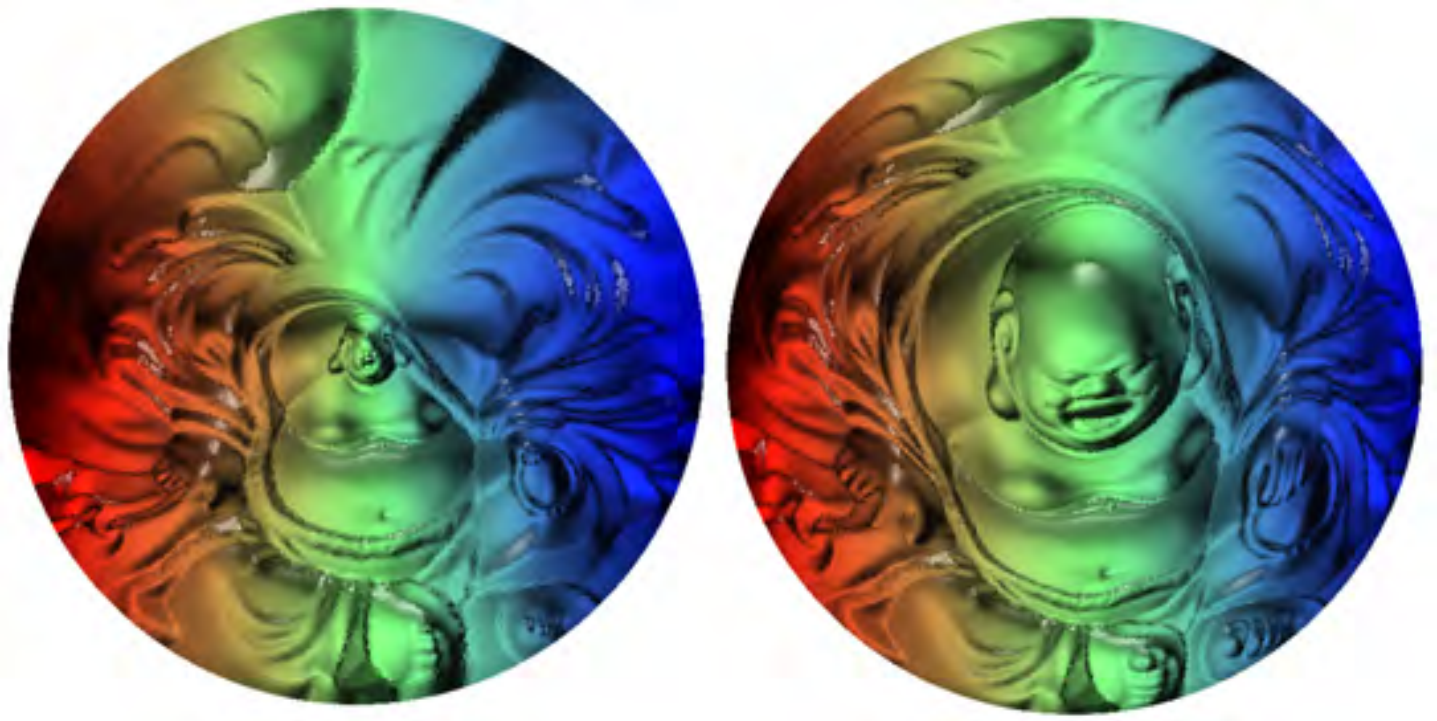
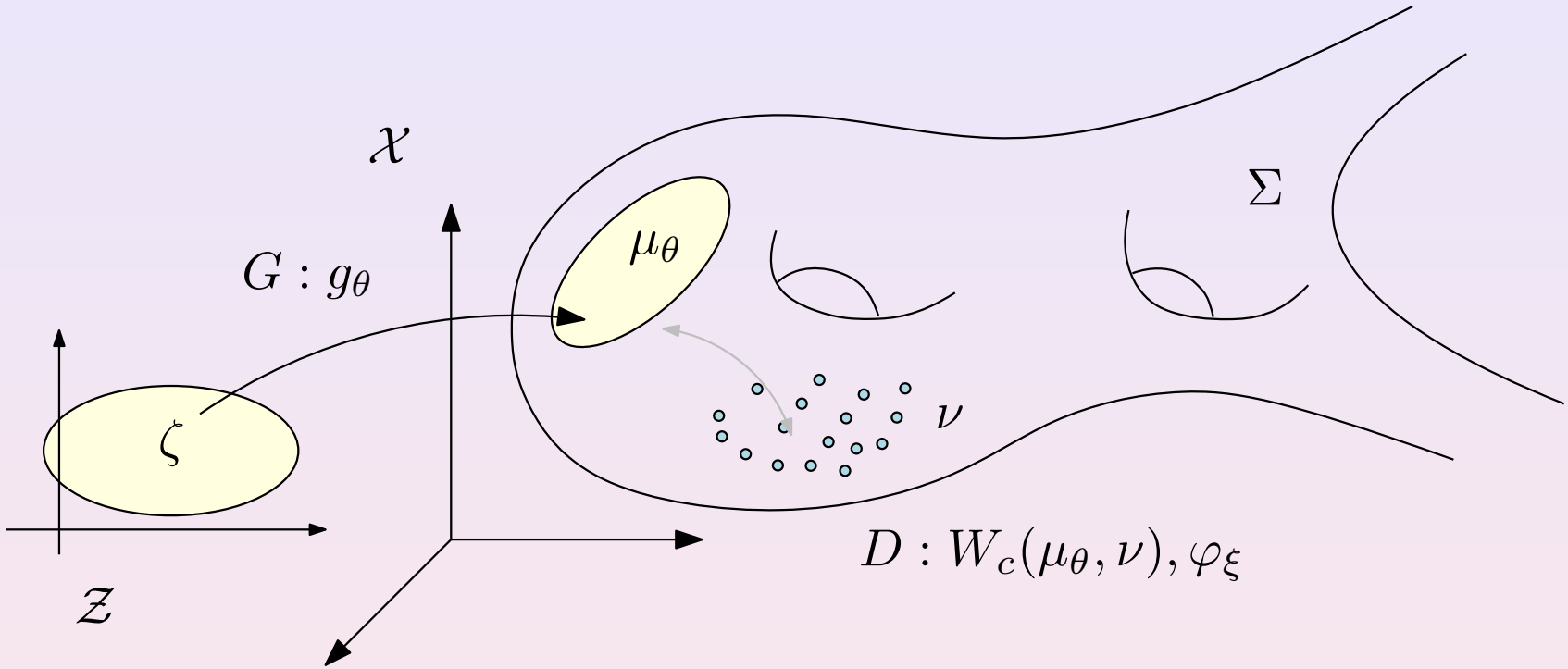


Figure: Optimal transport map between the conformal image and uniform distribution on the disk.

Wasserstein GAN Model



ν-training data distribution; ζ-uniform distribution;
$\mu_\theta = g_{\theta\#}\zeta$-generated distribution; G - generator computes g_θ;
D -discriminator, measures the distance between ν and μ_θ,
$W_c(\mu_\theta, \nu)$.

L^1 case

When $c(x,y) = |x-y|$, $\varphi^c = -\varphi$, given φ is 1-Lipsitz, the WGAN model: min-max optimization

$$\min_{\theta}\max_{\xi}\int_X \varphi_\xi \circ g_\theta(z)d\zeta(z) - \int_Y \varphi_\xi(y)d\nu(y).$$

namely

$$\min_{\theta}\max_{\xi}\ \mathbb{E}_{z\sim\zeta}(\varphi_\xi \circ g_\theta(z)) - \mathbb{E}_{y\sim\nu}(\varphi_\xi(y)).$$

with the constraint that φ_ξ is 1-Lipsitz.

OMT view of WGAN

L^2 case

The discriminator D computes the optimal transportation map from the generated distribution to the real distribution; the generator G computes the optimal transportation map from the white noise to the generated distribution, hence in theory

- The composition of the maps of G and D directly gives the desired transportation map from the white noise to the real distribution;
- The competition between D and G is unnecessary.
- G and D should collaborate by sharing the intermediate results to improve the efficiency.

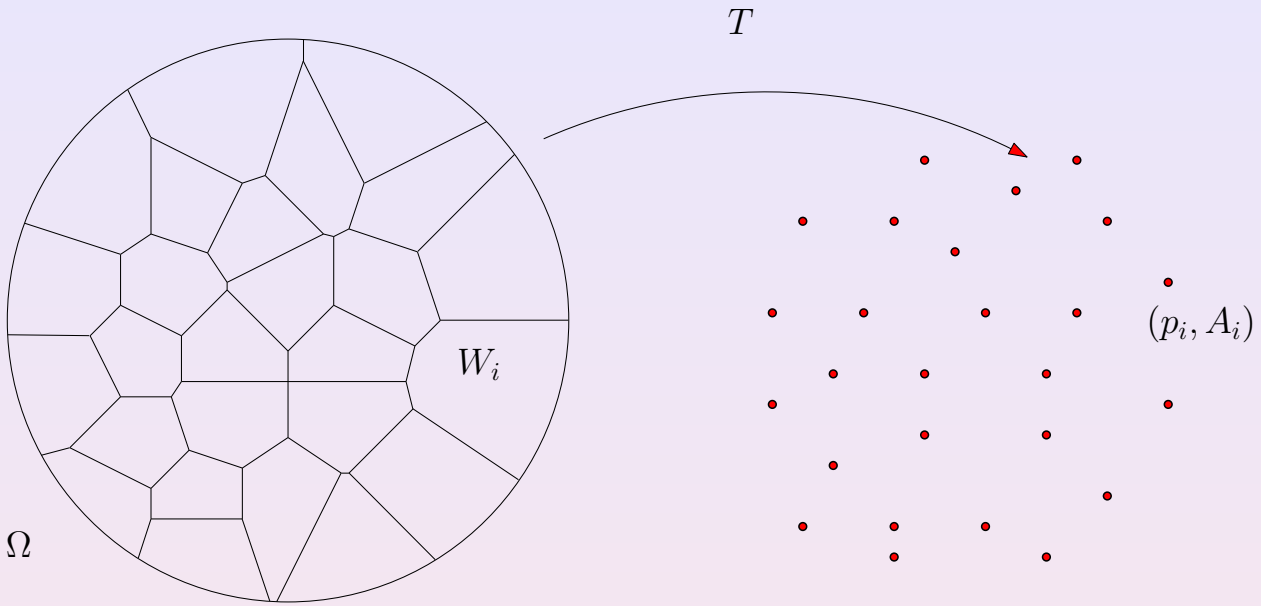
Empirical Distribution

In practice, the target probability measure is approximated by empirical distribution:

$$\nu = \sum_{i=1}^{n} \delta(y - y_i)\nu_i,$$

in general $\nu_i = 1/n$.

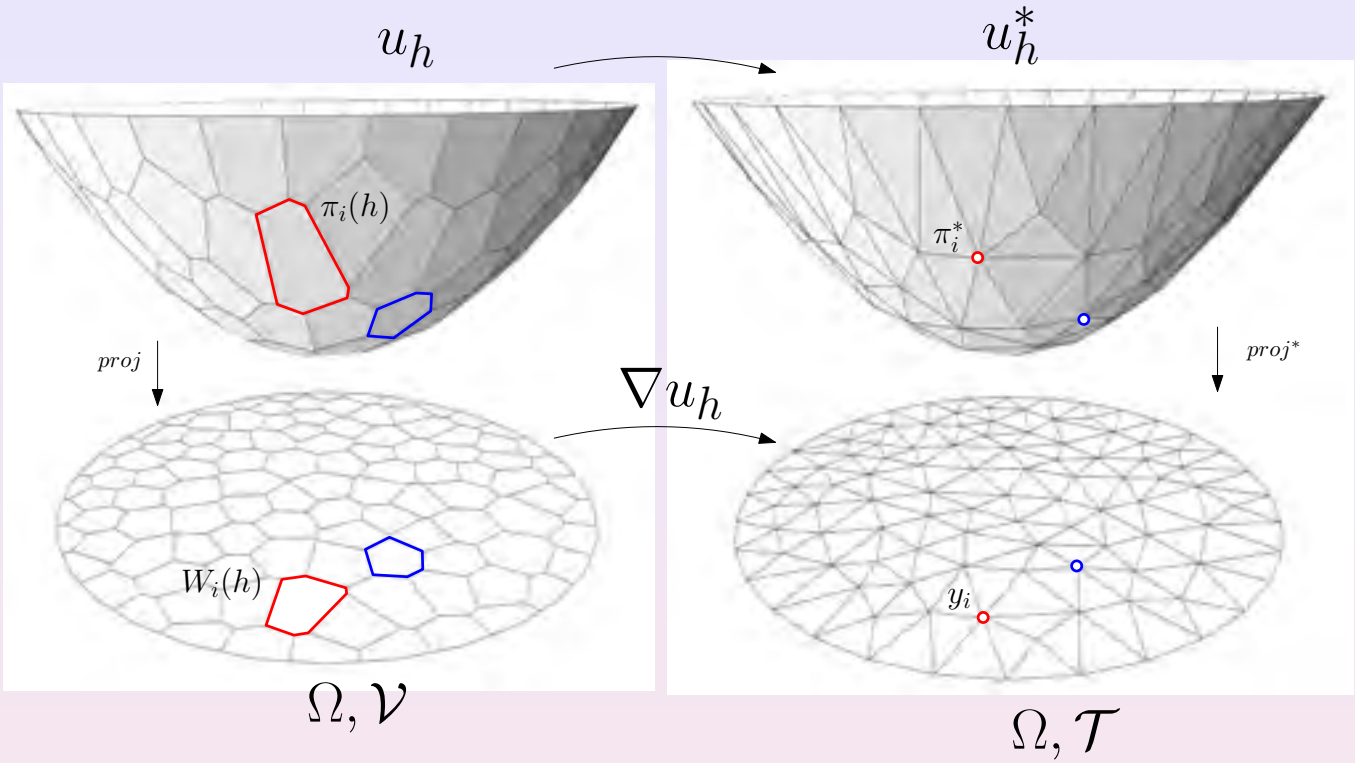
Semi-discrete Optimal Transportation



Given a compact convex domain Ω in $\mathbb{R}^n$ and $p_1, \cdots, p_k$ in $\mathbb{R}^n$ and $A_1, \cdots, A_k > 0$, find a transport map $T : U \to \{p_1, \cdots, p_k\}$ with $vol(T^{-1}(p_i)) = A_i$, so that T minimizes the transport cost

$$\frac{1}{2}\int_U |\mathbf{x} - T(\mathbf{x})|^2 d\mathbf{x}.$$

Power Diagram vs Optimal Transport Map



1. $\forall y_i \in Y$, construct a hyper-plane $\pi_h^i(x) = \langle x, y_i \rangle - h_i$;
2. compute the upper envelope of the planes $u_h(x) = \max_i \{\pi_h^i(x)\}$
3. produce the power diagram of Ω, $\mathcal{V}(h) = \cup_i W_i(h)$;
4. adjust the heights h, such that $\mu(W_i(h)) = \nu_i$.

Learning Problem

Problem

Since DNNs have large capacities, do they really learn anything or just memorize the training samples ?

Answer

The DL system learns the Brenier potential implicitly,

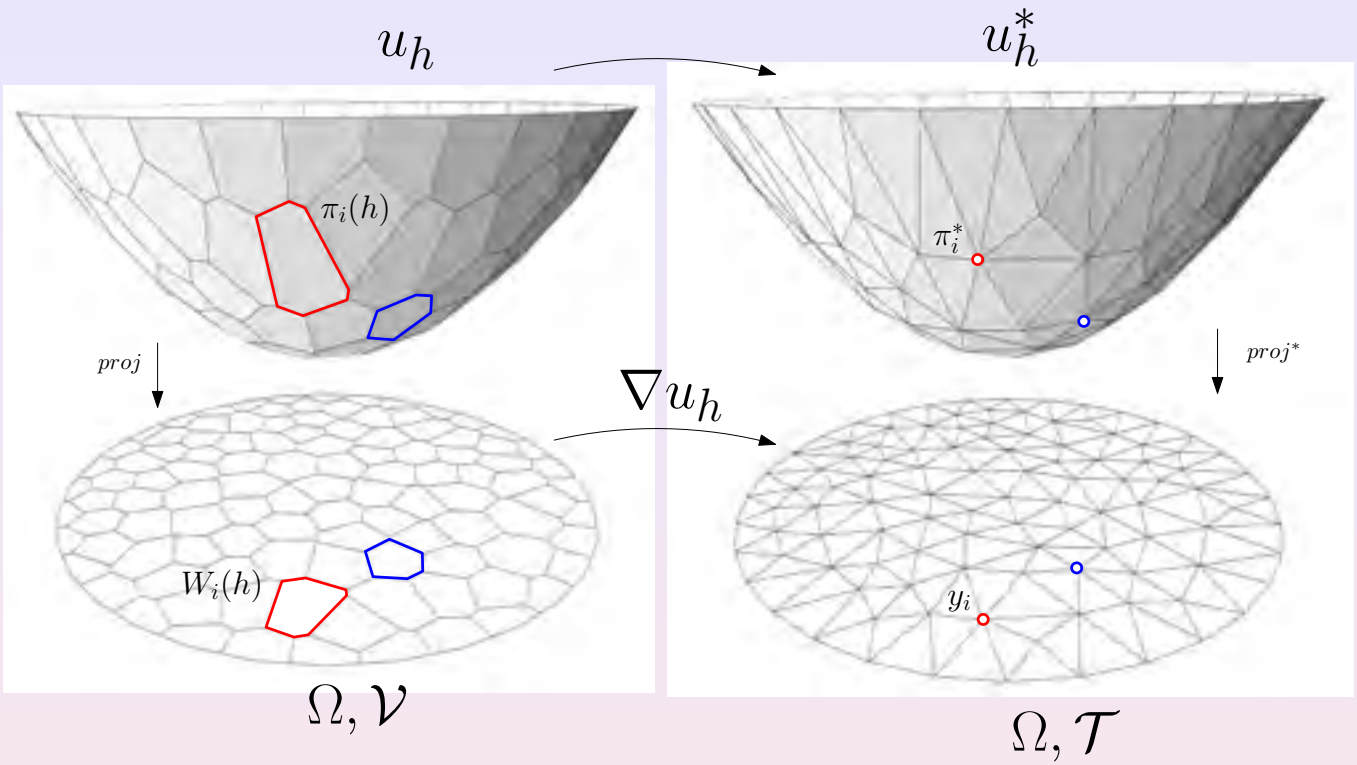
$$\max_i \{\langle x, y_i \rangle - h_i\}$$

It both learns and memorizes:

1. it memorizes all the training samples $\{y_i\}$;
2. it learns the probability $\{h_i\}$.

Complexity of Geometric Optimal Transport

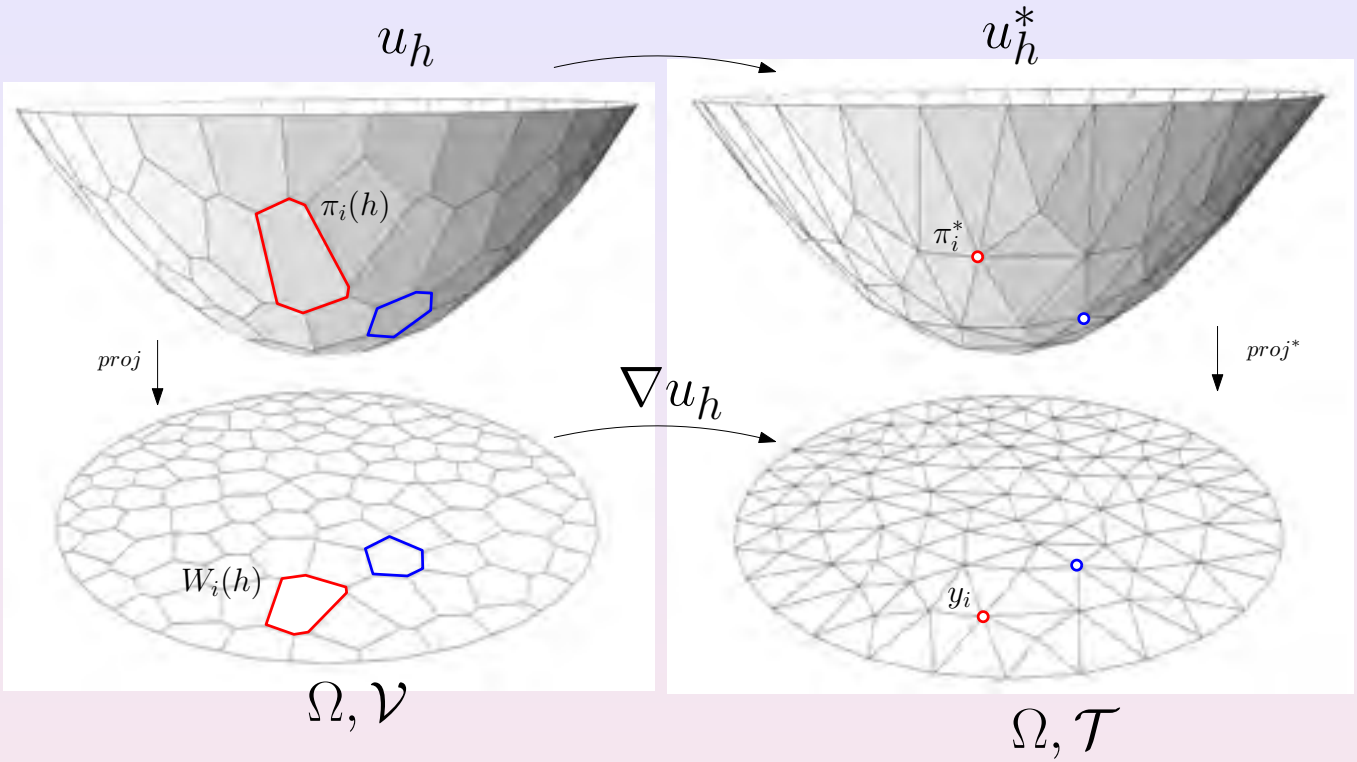
Semi-discrete Optimal Transport Map



A convex domain $\Omega \subset \mathbb{R}^d$ with a measure μ, $d\mu = f(x)dx$, $f(x)$ is continuous; the range is $Y = \{y_1, y_2, \ldots, y_n\}$ with measure $\mu = \sum_{i=1}^n \nu_i \delta(y - y_i)$. $\mu(\Omega) = \sum_{i=1}^n \nu_i$.

$$u_h := \max_{i=1}^{n} \{\langle x, y_i \rangle - h_i\}, \quad T = \nabla u_h.$$

Semi-discrete Optimal Transport Map

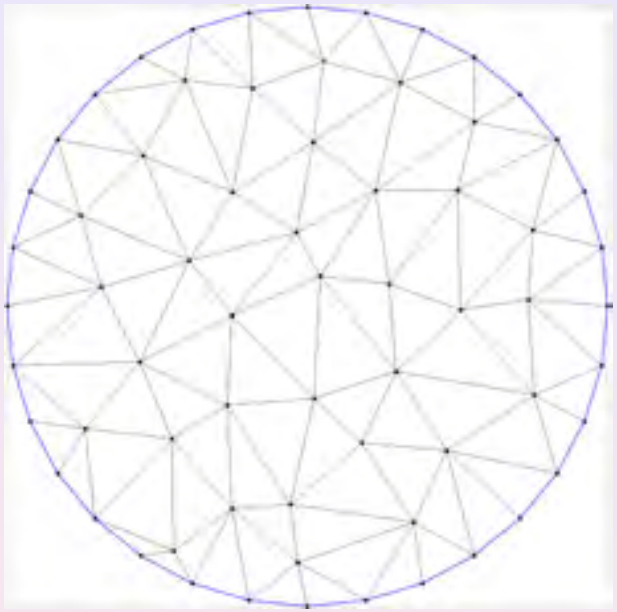


Power diagram and Convex polytope

$$\mathscr{D}_{\mathbf{h}} \quad \Omega = \bigcup_{i=1}^{n} W_i(\mathbf{h}), \quad W_i(\mathbf{h}) := \{x \in \mathbb{R}^d | \nabla u_h(x) = y_i\}$$

$$\mathscr{P}_{\mathbf{h}} \quad \text{Convex Hull}(\{(y_i, h)\}_{i=1}^{n}), \quad \text{graph of } u_h^*$$

Gelfand-Kapranov-Zelevinsky: Secondary Polytope



All the triangulations $\mathscr{T} \in \partial \mathscr{P}_Y$ are called *regular triangulations*.

Definition (Characteristics)

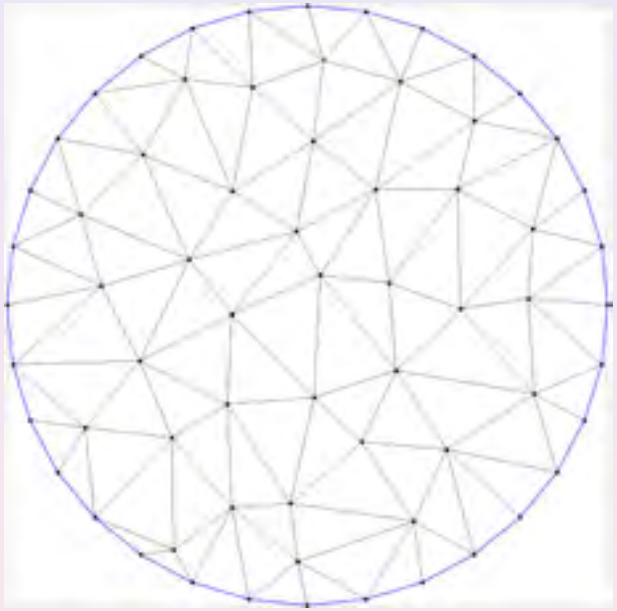
Let $\mathscr{T}$ is a triangulation of the convex hull of Y, Conv(Y). The characteristic of $\mathscr{T}$ defined as a vector, $c_{\mathscr{T}} = (a_1, a_2, \ldots, a_k)^T$,

$$a_i = \sum_{v_i \sim \sigma_j} \text{vol}(\sigma_j).$$

Definition (Secondary Polytop)

The secondary polytope of Y is the convex hull of all characteristic vectors of all possible triangulations of Conv(Y), denoted as $\mathscr{P}_Y$.

Gelfand-Kapranov-Zelevinsky: Secondary Polytope

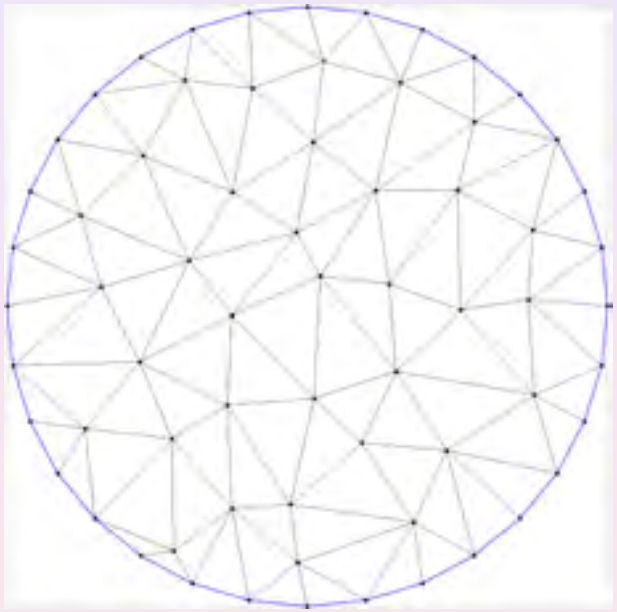


- Each upper envelope of planes

$$u_{\mathbf{h}} = \max_{i=1}^{n} \{\langle x, y_i \rangle - h_i\}$$

induces a closest power diagram $D_{\mathbf{h}}$. Its dual weighted Delaunay triangulation $\mathscr{T}_{\mathbf{h}}$ must be on the lower part of the secondary polytope $\mathscr{P}_Y$, and vice versa.

Gelfand-Kapranov-Zelevinsky: Secondary Polytope

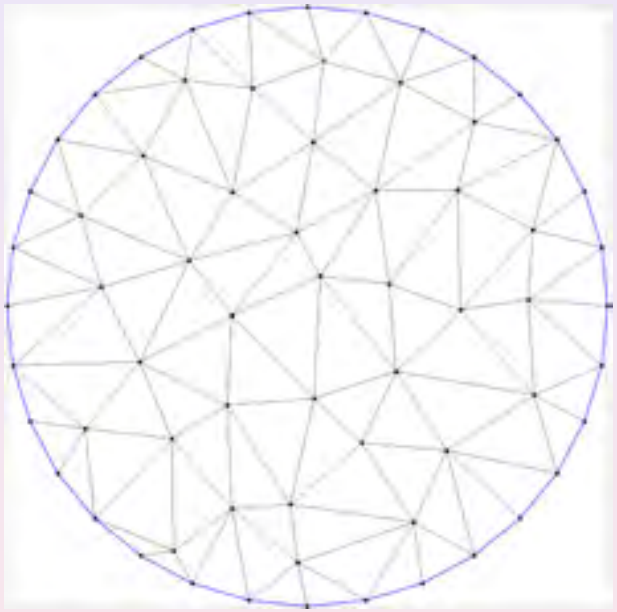


- Each lower envelop of planes

$$u_{\mathbf{h}} = \min_{i=1}^{n} \{\langle x, y_i \rangle - h_i\}$$

induces a furthest power diagram $D_{\mathbf{h}}$. Its dual weighted Delaunay triangulation $\mathscr{T}_{\mathbf{h}}$ must be on the upper part of the secondary polytope $\mathscr{P}_Y$, and vice versa.

Gelfand-Kapranov-Zelevinsky: Secondary Polytope



- The number of regular triangulations of any configuration of n points in dimension d is $O\left(n^{(n-d-2)^2}\right)$.
- The diameter of the secondary polytope is bounded by

$$\min\left\{(d+2)\binom{n}{\lfloor \frac{d}{2}+1 \rfloor}, \binom{n}{d+2}\right\}$$

Secondary Diagram

Definition (Admissible Height Space)

Fix Ω and Y, the admissible height space is defined as

$$\mathscr{H}_Y = \{\mathbf{h} | W_i(\mathbf{h}) \neq \emptyset, \quad \forall W_i(\mathbf{h}) \in \mathscr{D}_{\mathbf{h}}\} \bigcap \left\{ \sum_{i=1}^{n} h_i = 0 \right\}.$$

Theorem (Gu-Luo-Sun-Yau)

The admissible height space $\mathscr{H}_Y$ is a convex non-empty set.

Proof.

By Alexandrov theorem and Brunn-Minkowski inequality. □

Secondary Diagram

Definition (Secondary Diagram)

Fix Ω and Y, the admissible height space $\mathscr{H}_Y$ has a cell decomposition:

$$\mathscr{D}_Y \quad \mathscr{H}_Y := \bigcup_{\mathscr{T} \in \mathscr{P}_Y} H_Y(\mathscr{T}), \quad H_Y(\mathscr{T}) := \{\mathbf{h} |\ u^*_{\mathbf{h}} \text{ induces } \mathscr{T}\}$$

Theorem (Gu-Lei-Si)

The secondary diagram $\mathscr{D}_Y$ is a power diagram, induced by lower envelop of hyperplanes

$$\min \{\langle \mathbf{h}, \mathbf{c}_{\mathscr{T}} \rangle | \mathscr{T} \in \mathscr{P}_Y\}.$$

Singularity of Geometric Optimal Transport Map

Regularity of Optimal Transportation Map

Let Ω and Ω^* be bounded domains in $\mathbb{R}^n$, let f and g be mass densities on Ω and Ω^* satisfying

1. $0 \leq f \in L^1(\Omega)$, $0 \leq g \in L^1(\Omega^*)$,

$$\int_\Omega f = \int_{\Omega^*} g.$$

2. $\exists$ constants $f_0, f_1, g_0, g_1 > 0$, such that

$$f_0 \leq f \leq f_1, g_0 \leq g \leq g_1.$$

Regularity of Optimal Transportation Map

Let (u, v) be the Kantorovich's potential functions. The optimal mapping T_u is given by

$$Du(x) = D_x c(x, T_u(x))$$

Differentiate the formula

$$D^2 u(x) = D_x^2 c(x, T_u(x)) + D_{xy}^2 c(x, T_u(x)) DT_u.$$

We obtain the equation

$$\det[D^2 u(x) - D_x^2 c(x, T_u(x))] = \det D_{xy}^2 c(x, T_u(x)) \frac{f(x)}{g(T_u(x))},$$

with the boundary condition $T_u(\Omega) = \Omega^*$.

Regularity of Optimal Transportation Map

Caffarelli obtained the regularity of optimal mappings for the cost function

$$c(x,y) = |x-y|^2$$

or equivalently $c(x,y) = x \cdot y$, then we have the standard Monge-Ampere equation

$$\det D^2 u = \frac{f(x)}{g(Du(x))},$$

with boundary condition $Du(\Omega) = \Omega^*$.

1. if $f, g > 0, \in C^\alpha$ and Ω^* is convex, then $u \in C^{2,\alpha}(\Omega)$
2. if $f, g > 0, \in C^0$ and Ω^* is convex, then $u \in W^{2,p}_{loc}(\Omega)$, $\forall p > 1$ (the continuity is needed for large p).
3. if $f, g > 0, \in C^\alpha$, both Ω and Ω^* are uniformly convex and $C^{2,\alpha}$, then $u \in C^{2,\alpha}(\bar{\Omega})$

Regularity of Optimal Transportation Map

Theorem (Ma-Trudinger-Wang)

The potential function u is C^3 smooth if the cost function c is smooth, f, g are positive, $f \in C^2(\Omega)$, $g \in C^2(\Omega^*)$, and

- *A1* $\forall x, \xi \in \mathbb{R}^n$, $\exists! y \in \mathbb{R}^n$, *s.t.* $\xi = D_x c(x, y)$ *(for existence)*
- *A2* $|D^2_{xy} c| \neq 0$.
- *A3* $\exists c_0 > 0$ *s.t.* $\forall \xi, \eta \in \mathbb{R}^n$, $\xi \perp \eta$

$$\sum (c_{ij,rs} - c^{p,q} c_{ij,p} c_{q,rs}) c^{r,k} c^{s,l} \xi_i \xi_j \eta_k \eta_l \geq c_0 |\xi|^2 |\eta|^2.$$

- *B1* Ω^* *is c-convex w.r.t.* Ω, *namely* $\forall x_0 \in \Omega$,

$$\Omega^*_{x_0} := D_x c(x_0, \Omega^*)$$

is convex.

Optimal Transportation Map

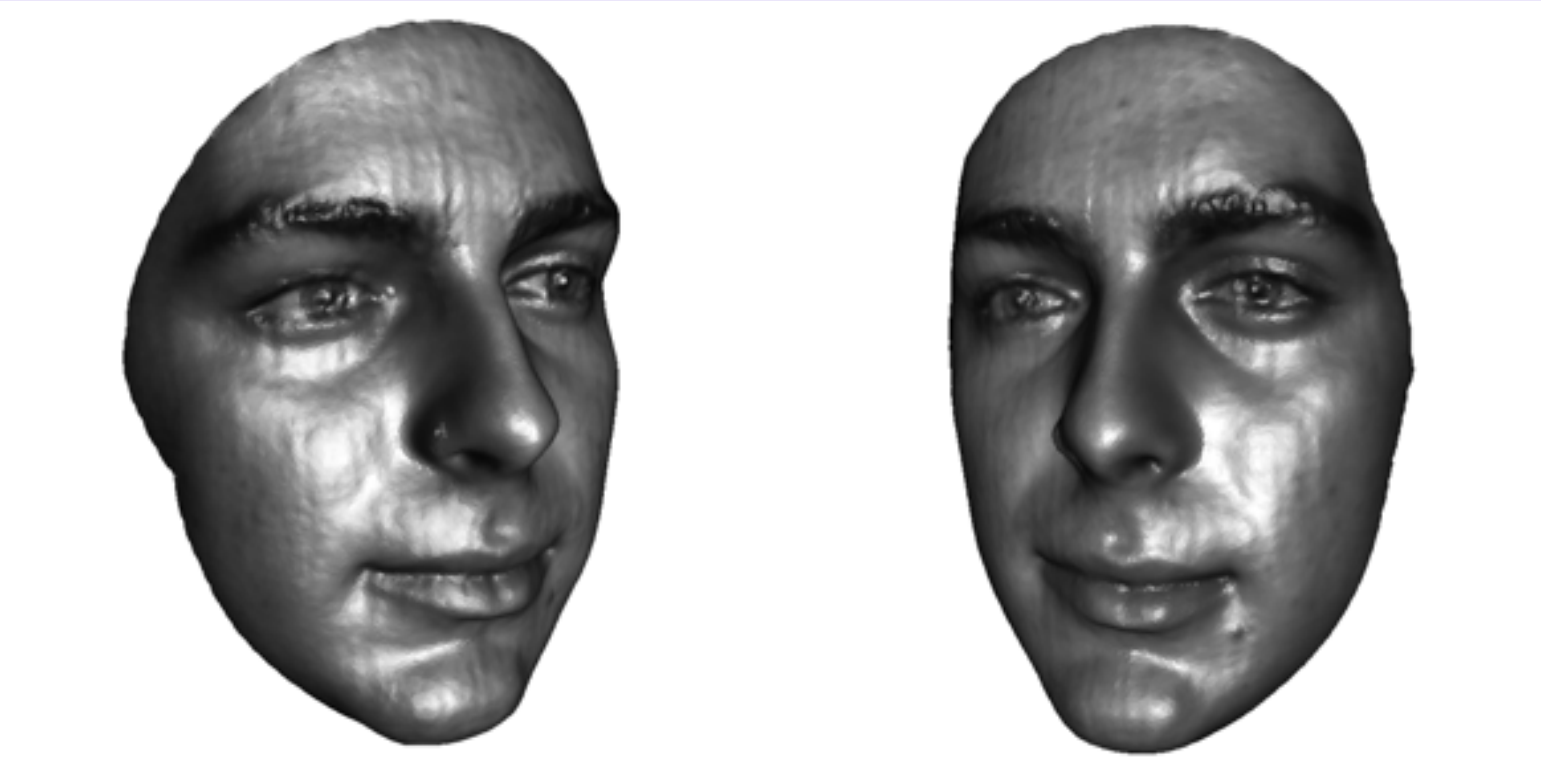


Figure: Optimal transportation map.

Optimal Transportation Map

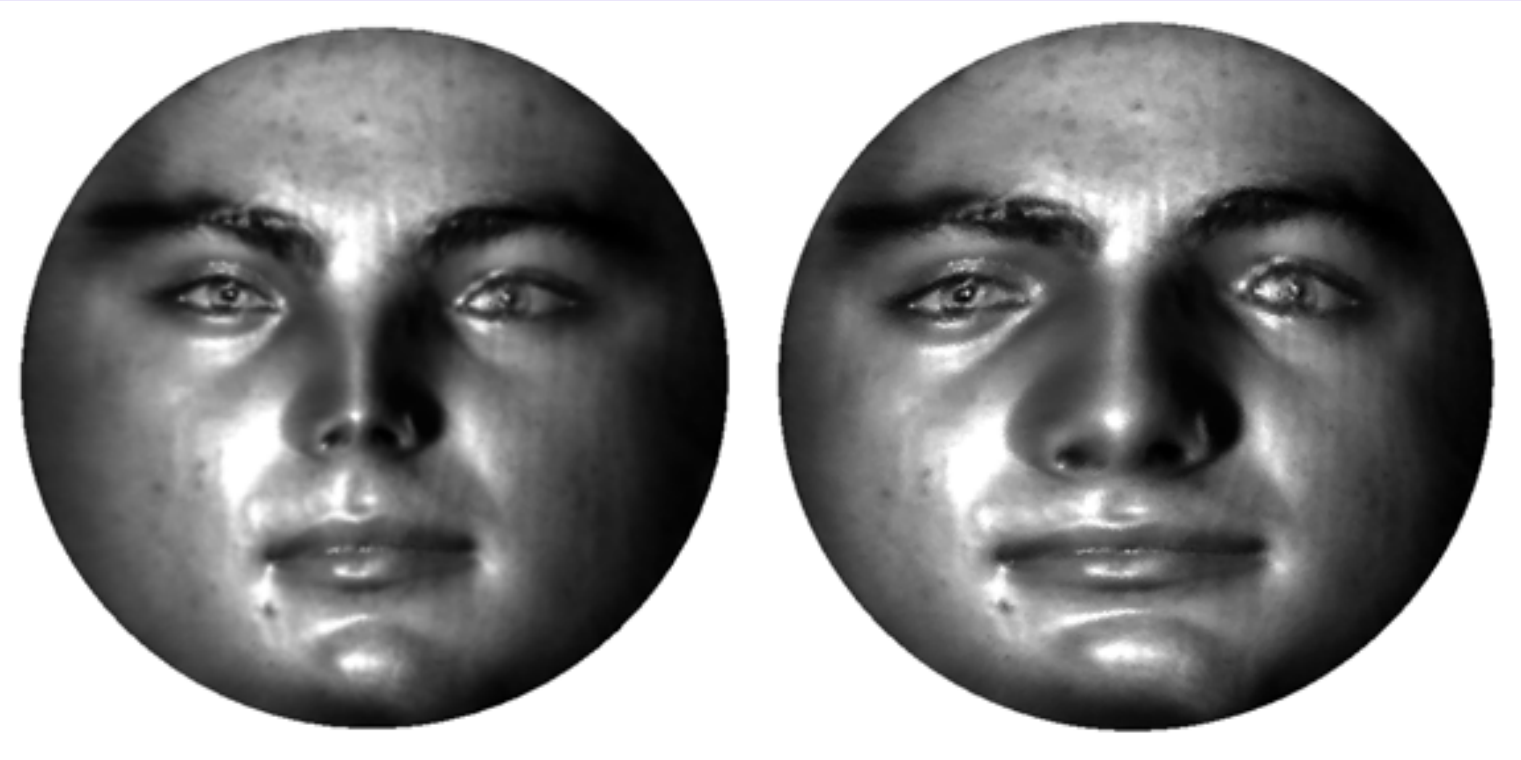


Figure: Optimal transportation map.

Optimal Transportation Map

Figure: Optimal transportation map.

Theorem (Figalli Regularity)

Let $\Omega, \Lambda \subset \mathbb{R}^d$ be two bounded open sets, let $f, g : \mathbb{R}^d \to \mathbb{R}^+$ be two probability densities, that are zero outside Ω, Λ and are bounded away from zero and infinity on Ω, Λ, respectively. Denote by $T = \nabla u : \Omega \to \Lambda$ the optimal transport map provided by Brenier theorem. Then there exist two relatively closed sets $\Sigma_\Omega \subset \Omega$ and $\Sigma_\Lambda \subset \Lambda$ with $|\Sigma_\Omega| = |\Sigma_\Lambda| = 0$ such that $T : \Omega \setminus \Sigma_\Omega \to \Lambda \setminus \Sigma_\Lambda$ is a homeomorphism of class $C^{0,\alpha}_{loc}$ for some $\alpha > 0$.

Singularity Set of OT Maps

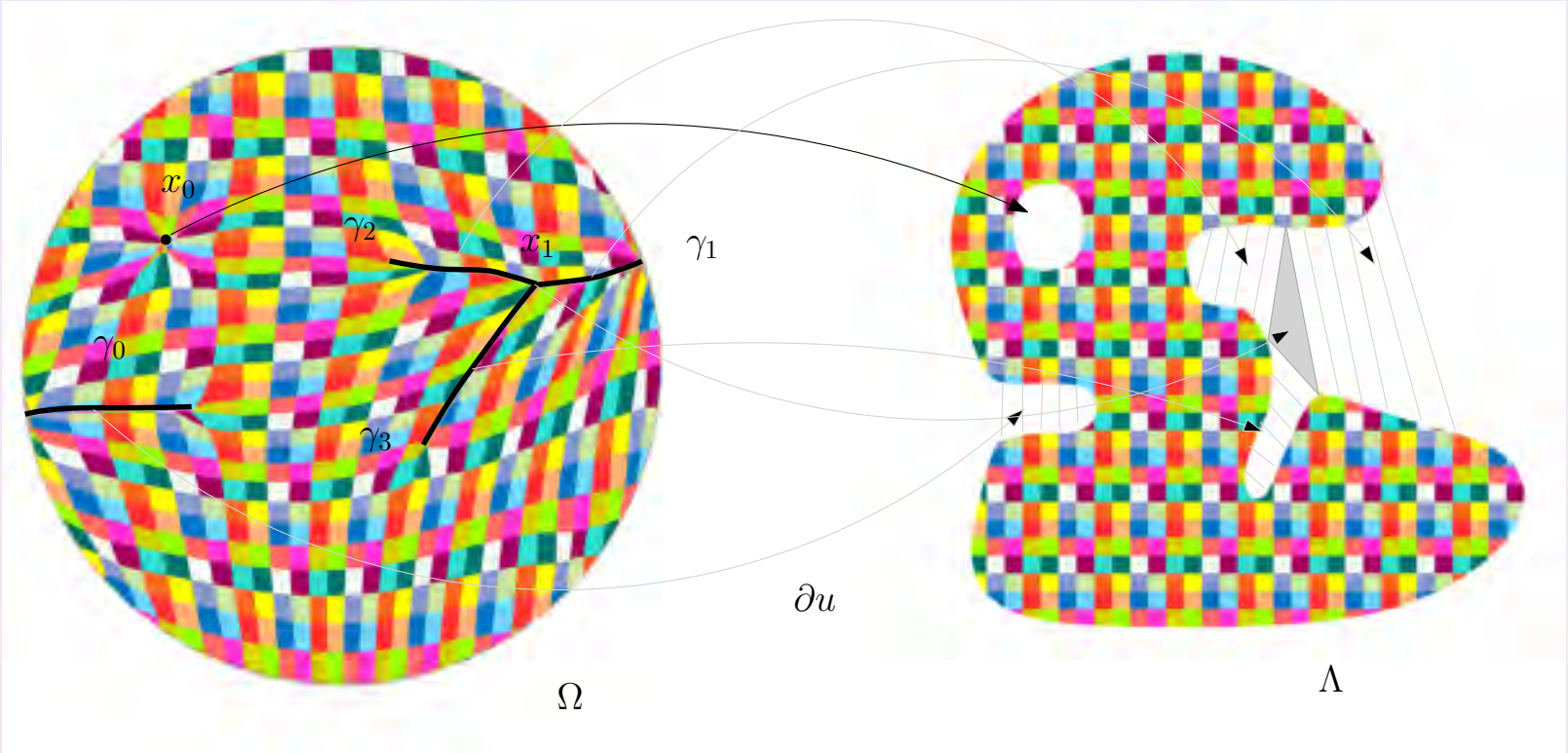
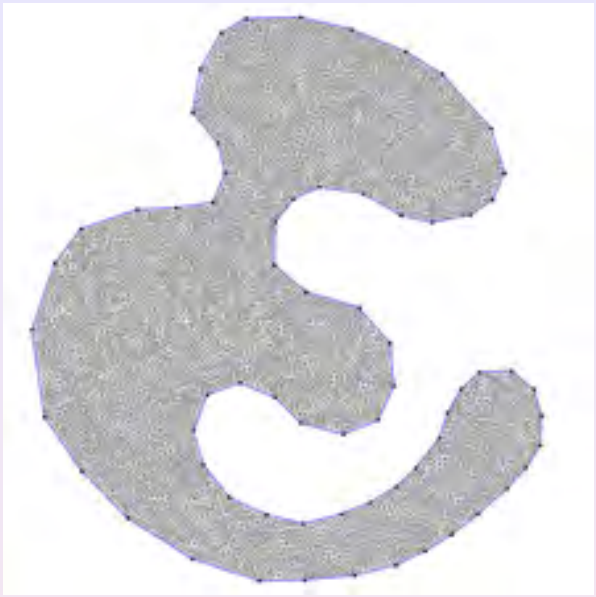
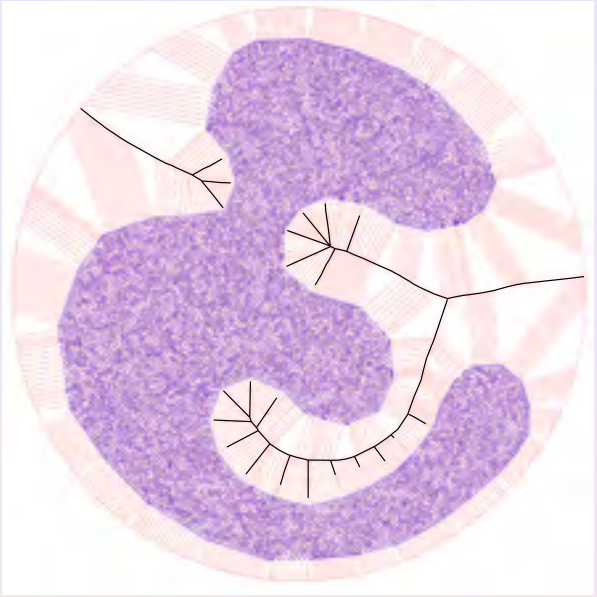


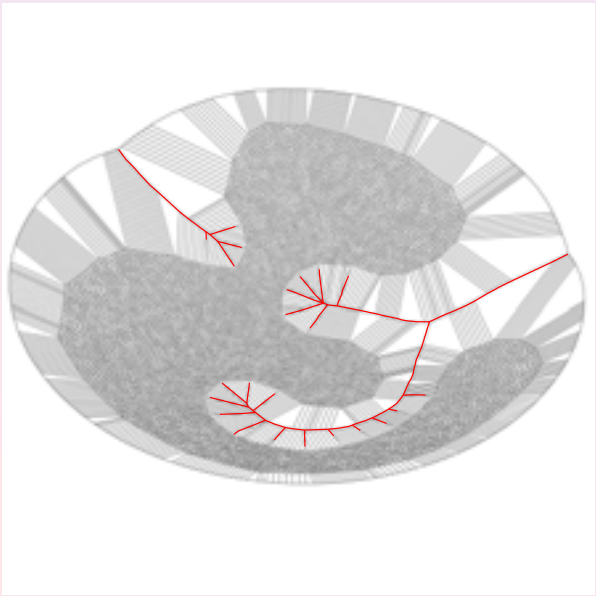
Figure: Singularity structure of an optimal transportation map.

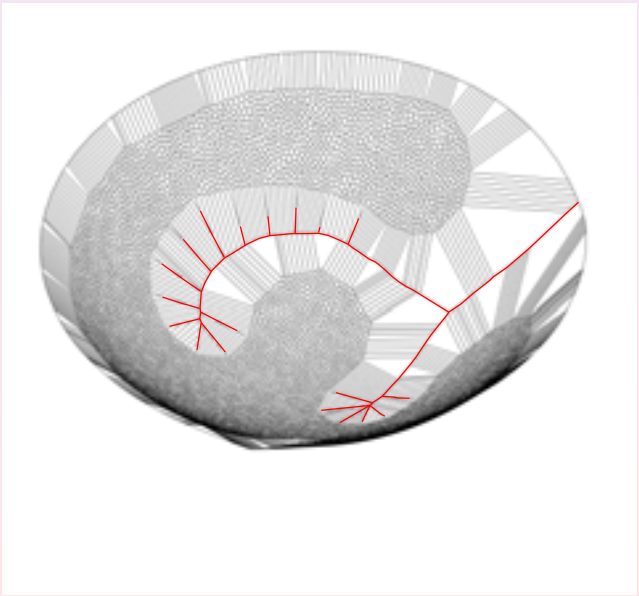
We call Σ_Ω as singular set of the optimal transportation map $\nabla u : \Omega \to \Lambda$.

Medial Axis

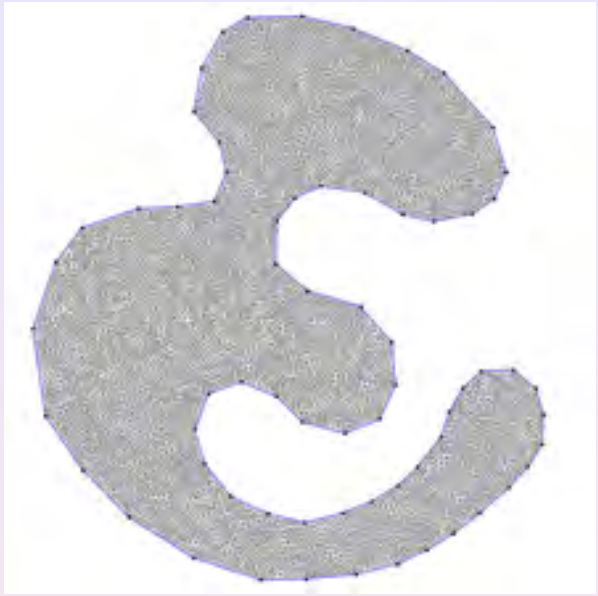


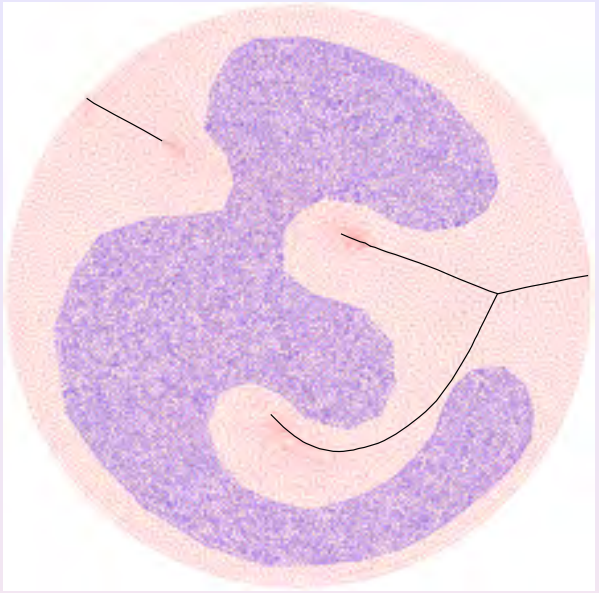


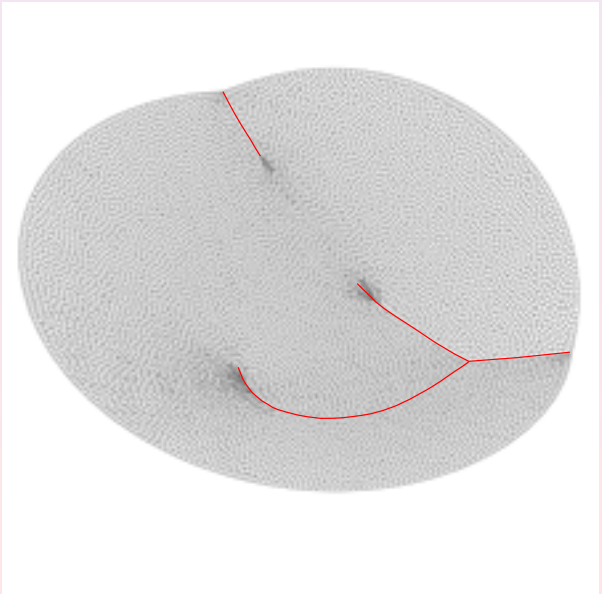




Singularity of OT Map







Given a planar polygonal domain Ω, we densely sample the boundary and the interior, the samples are denoted as $Y = \{y_1, y_2, \cdots, y_n\}$. Given the powers $\{w_1, w_2, \cdots, w_n\}$, or equivalently the height $\mathbf{h} = (h_1, h_2, \cdots, h_n)$, the power diagram is denoted as $\mathscr{D}_Y(\mathbf{h})$. For each point $p \in \mathbb{R}^2$, the *closest point* of p to Y is defined as

$$\mathrm{Cl}_Y(p, \mathbf{h}) := \mathrm{argmin}_i \mathrm{pow}(p, y_i)$$

Definition (Power Medial Axis)

Given $(Y, \mathbf{h})$, the power medial axis is defined as

$$\mathbf{MAT}_Y(\mathbf{h}) := \{p \in \mathbb{R}^2 | \ |\mathrm{Cl}_Y(p, \mathbf{h})| > 1\}.$$

Singularity Stability

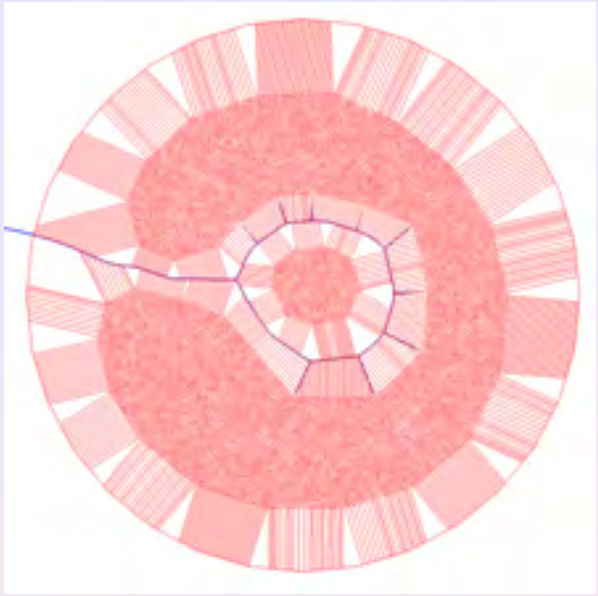
Theorem

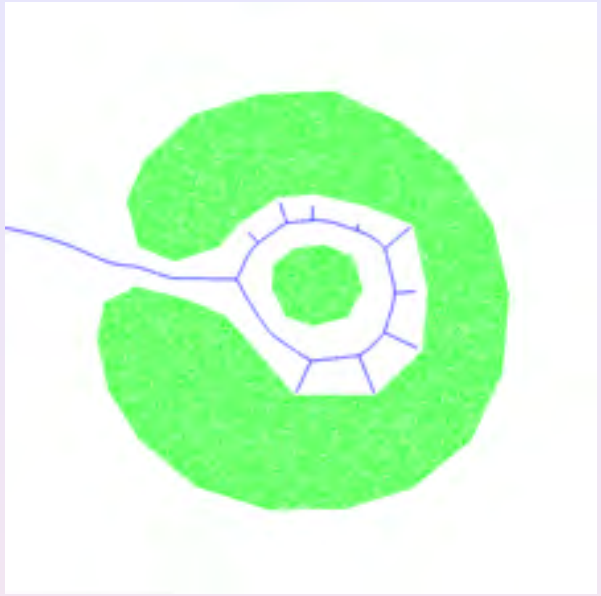
Given a convex domain $D \subset \mathbb{R}^n$ and the discrete point set $Y = \{y_1, y_2, \cdots, y_k\}$, then for any two admissible heights $\mathbf{h}_1, \mathbf{h}_2 \in \mathscr{H}_Y$, their power medial axises are homotopic to each other.

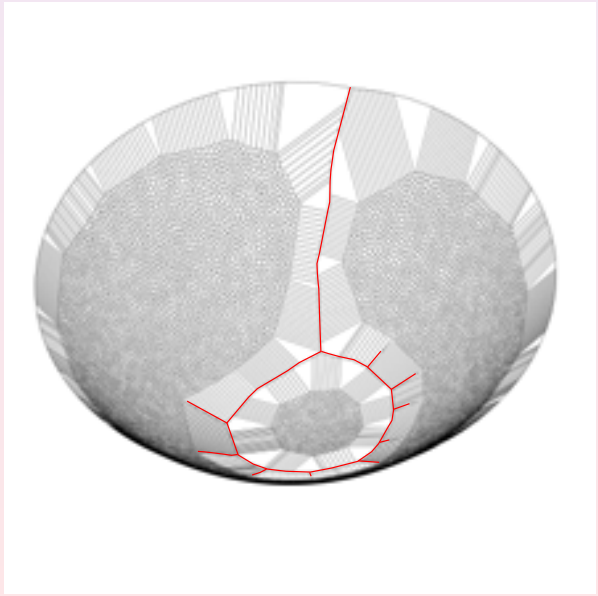
Corollary

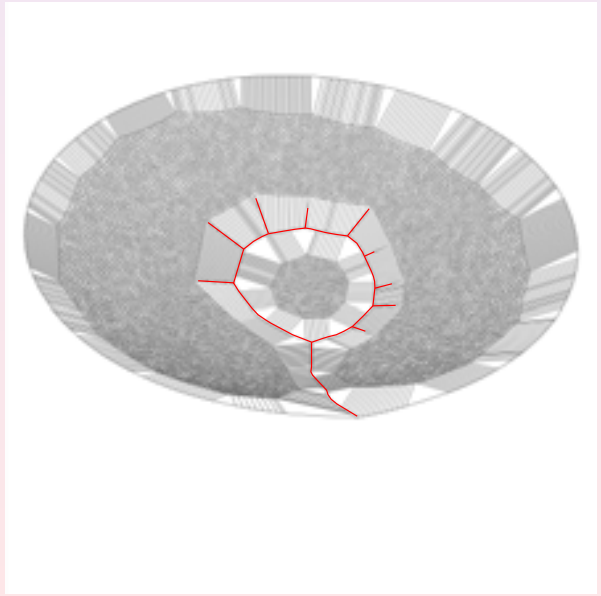
The singularity of a semi-discrete optimal transport map is homotopic to the medial axis.

singularity

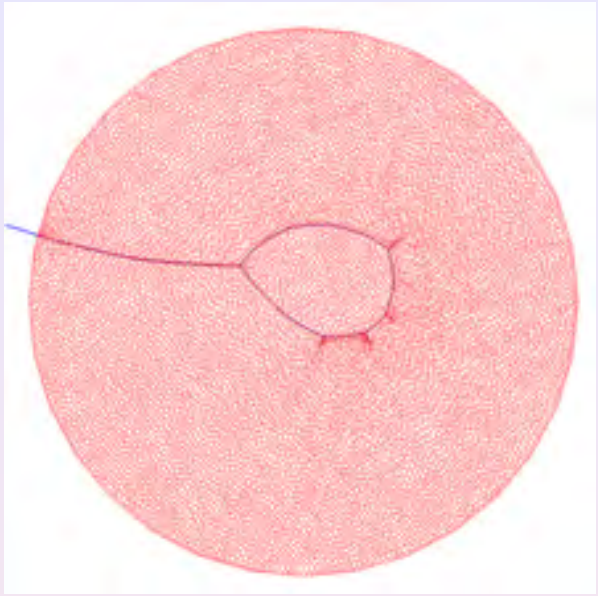


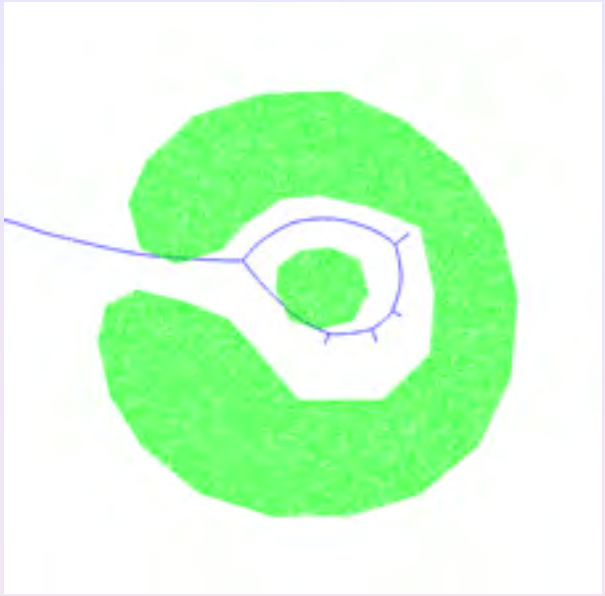


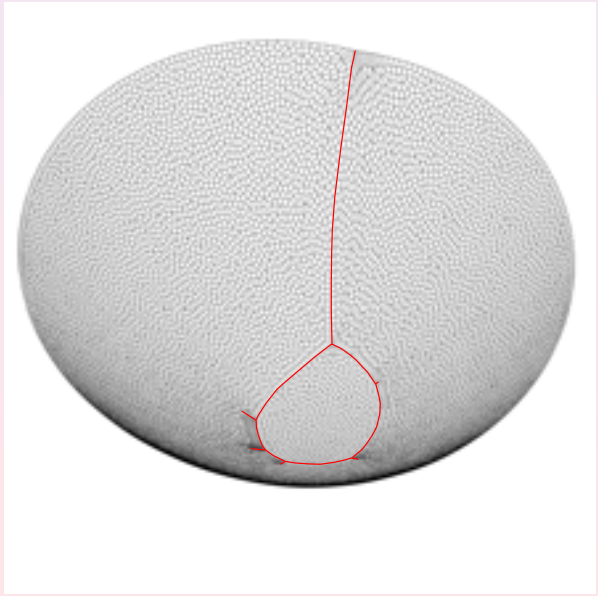


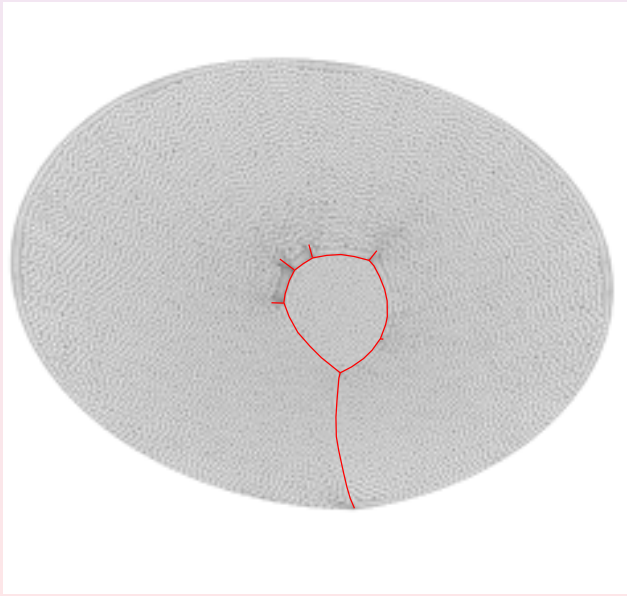


singularity

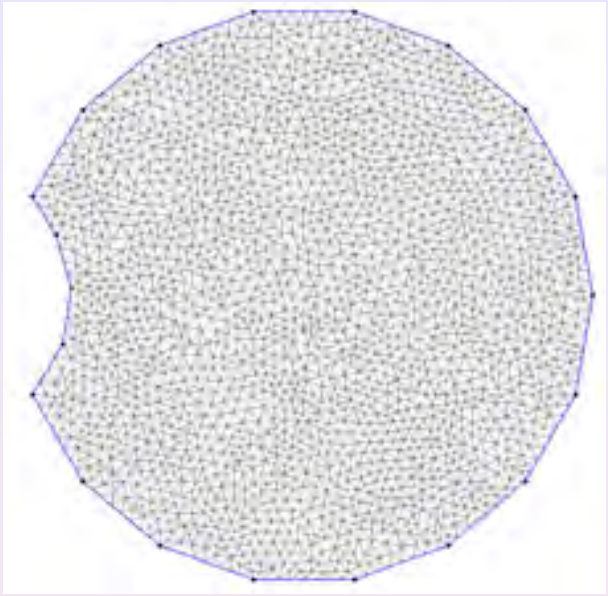


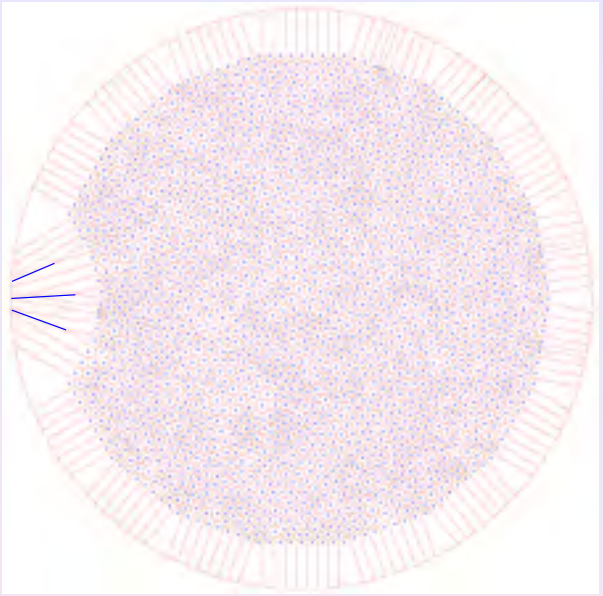


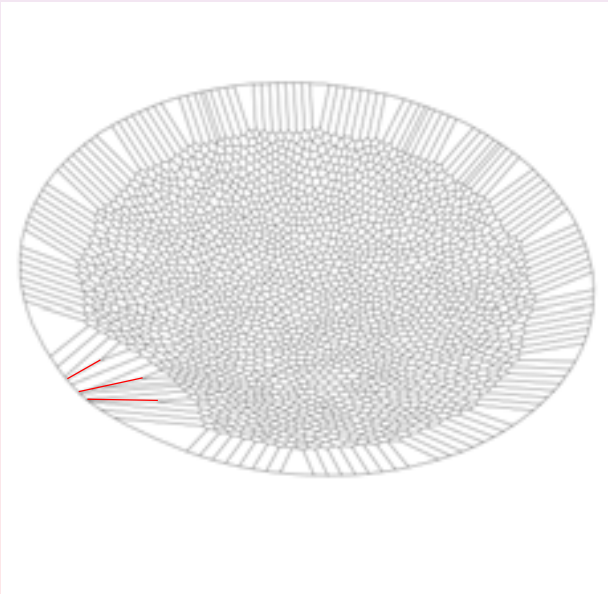




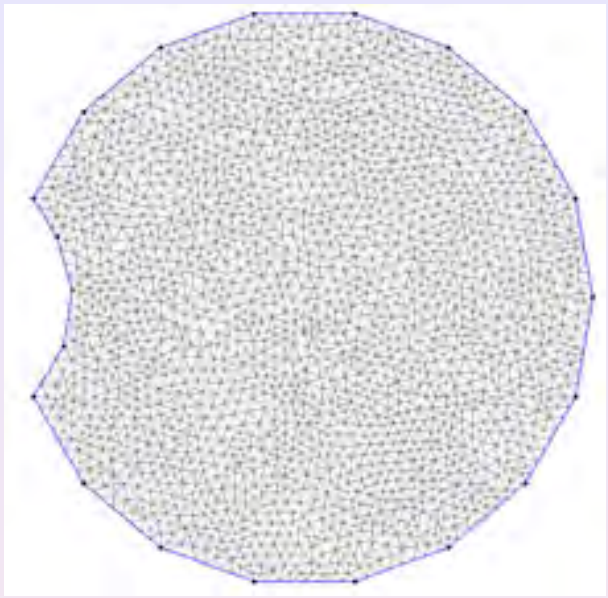
Singularity Stability



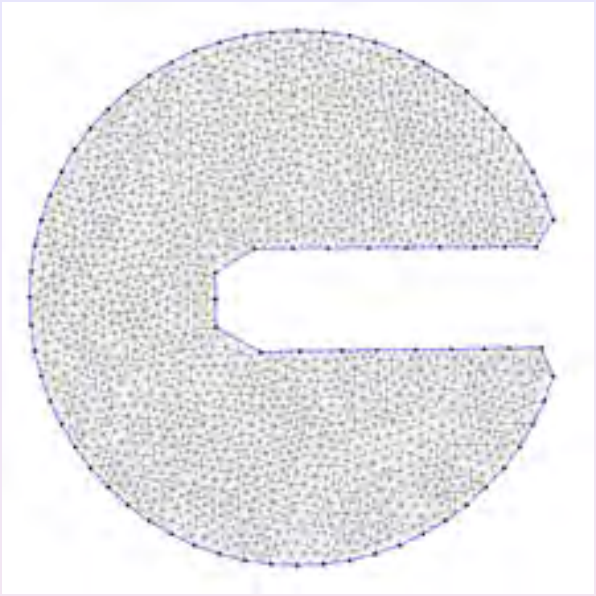


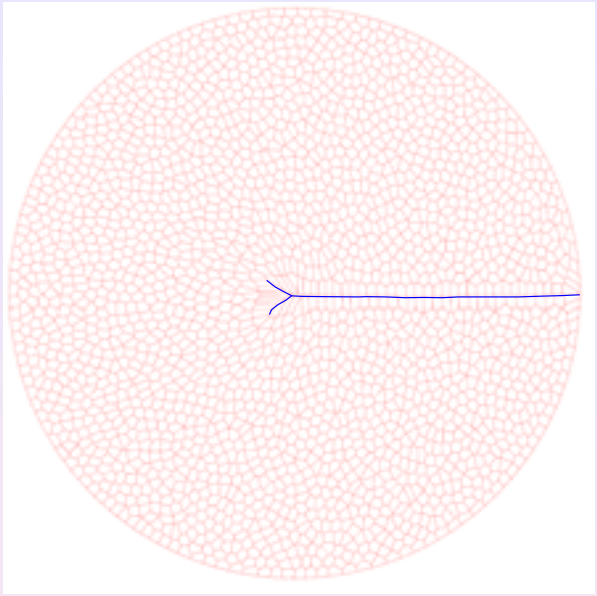


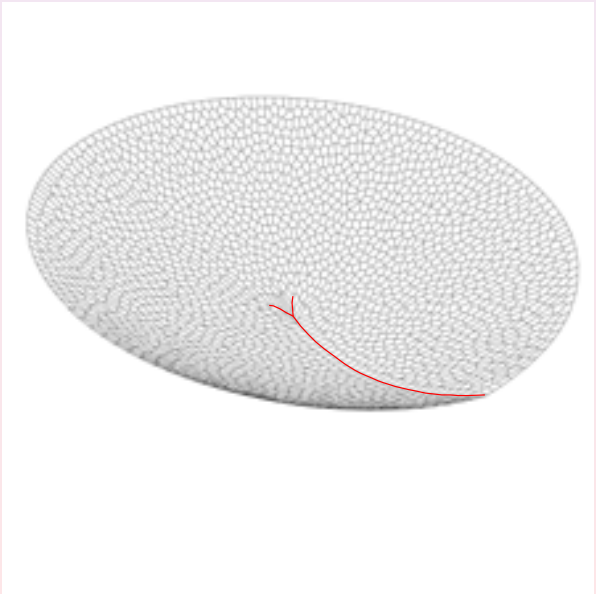
Singularity Stability



Singularity Stability







Discontinuity of Optimal Transportation Map

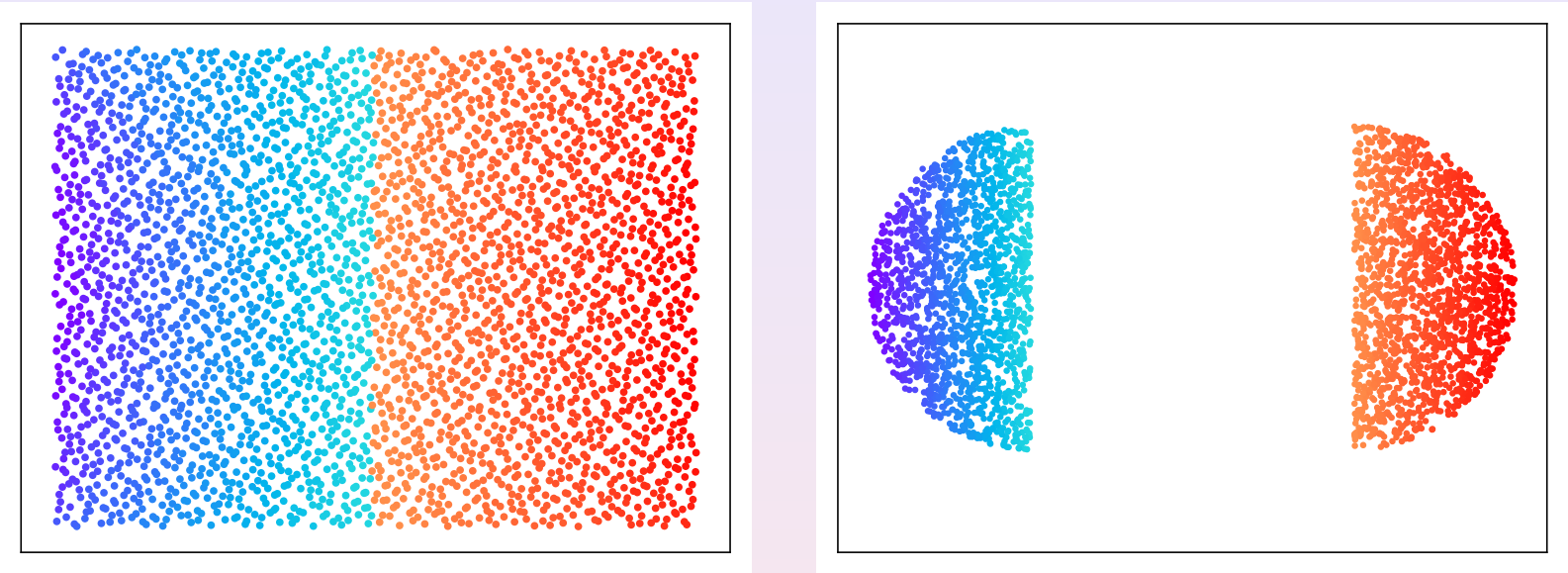


Figure: Discontinuous Optimal transportation map, produced by a GPU implementation of algorithm based on our theorem. The middle line is the singularity set Σ_1.

Discontinuity of Optimal Transportation Map

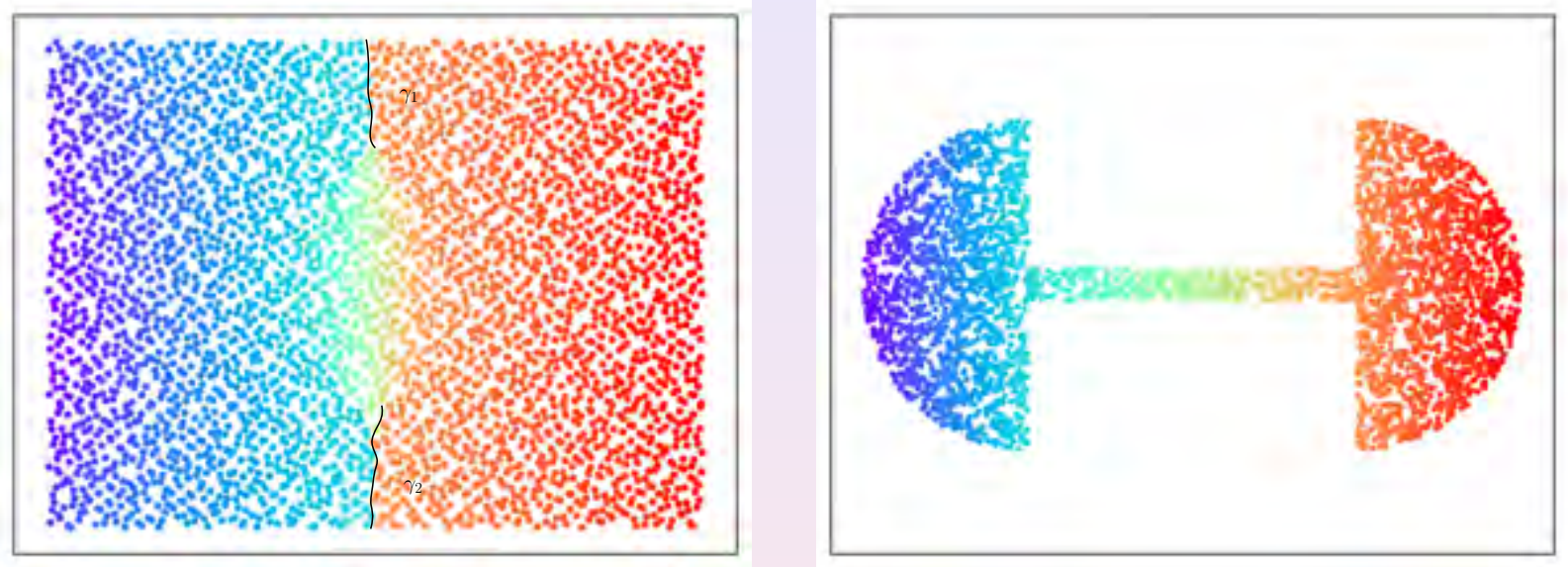


Figure: Discontinuous Optimal transportation map, produced by a GPU implementation of algorithm based on regularity theorem. γ_1 and γ_2 are two singularity sets.

Discontinuity of Optimal Transportation Map

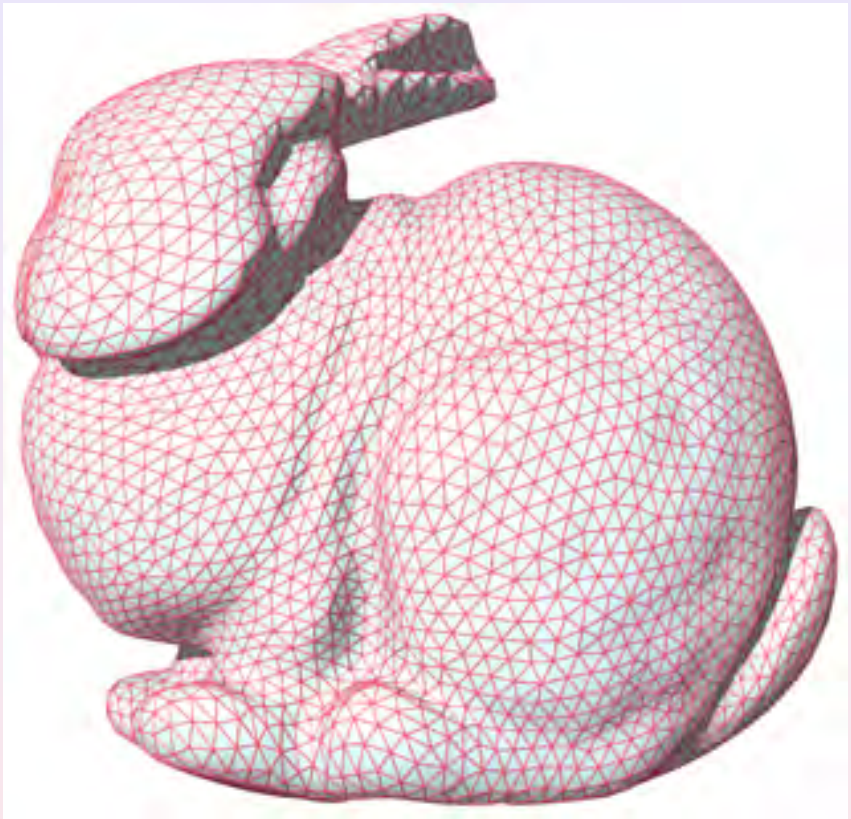


Figure: Optimal transportation between a solid ball to the Stanford bunny. The singular sets are the foldings on the boundary surface.

Discontinuity of Optimal Transportation Map

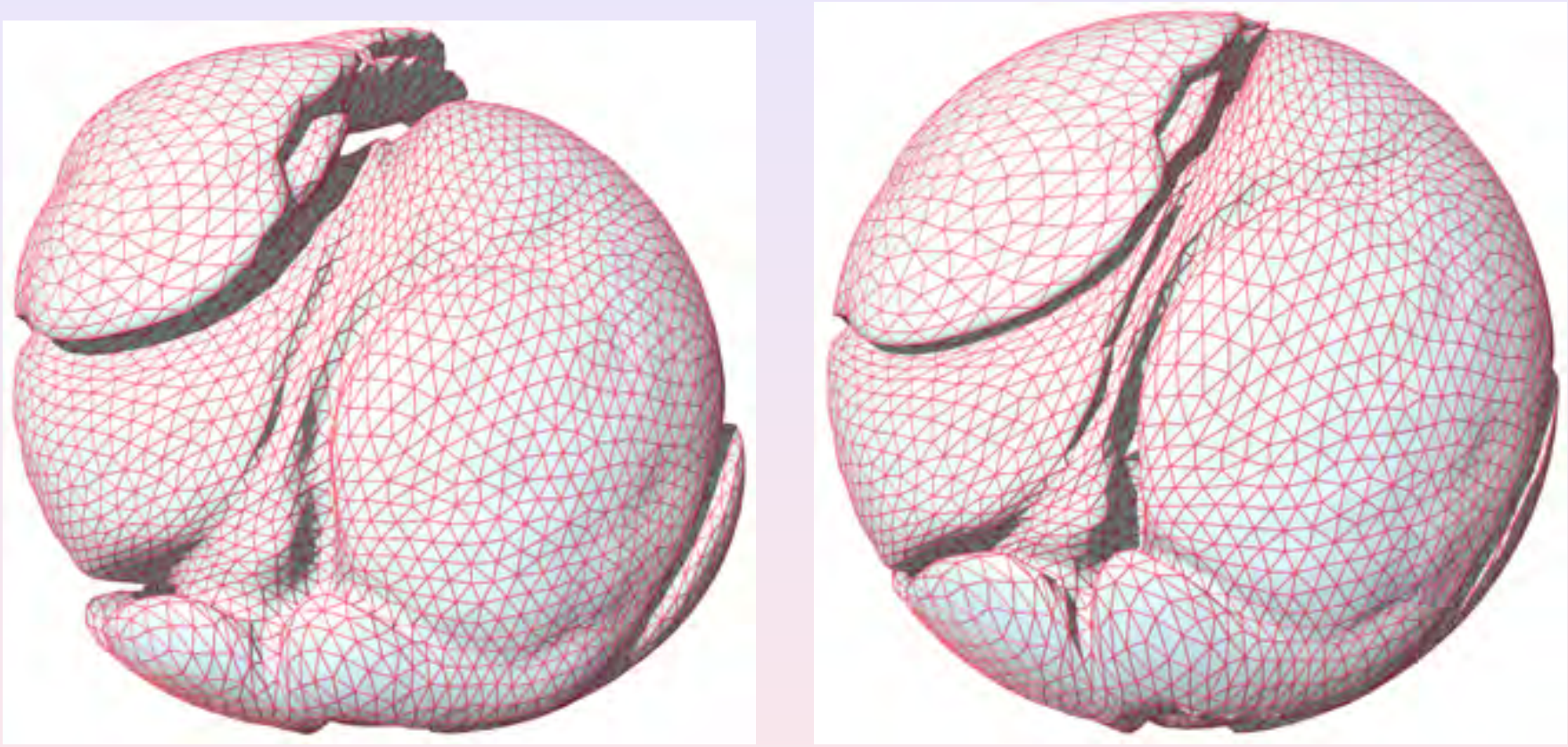


Figure: Optimal transportation between a solid ball to the Stanford bunny. The singular sets are the foldings on the boundary surface.

Discontinuity of Optimal Transportation Map

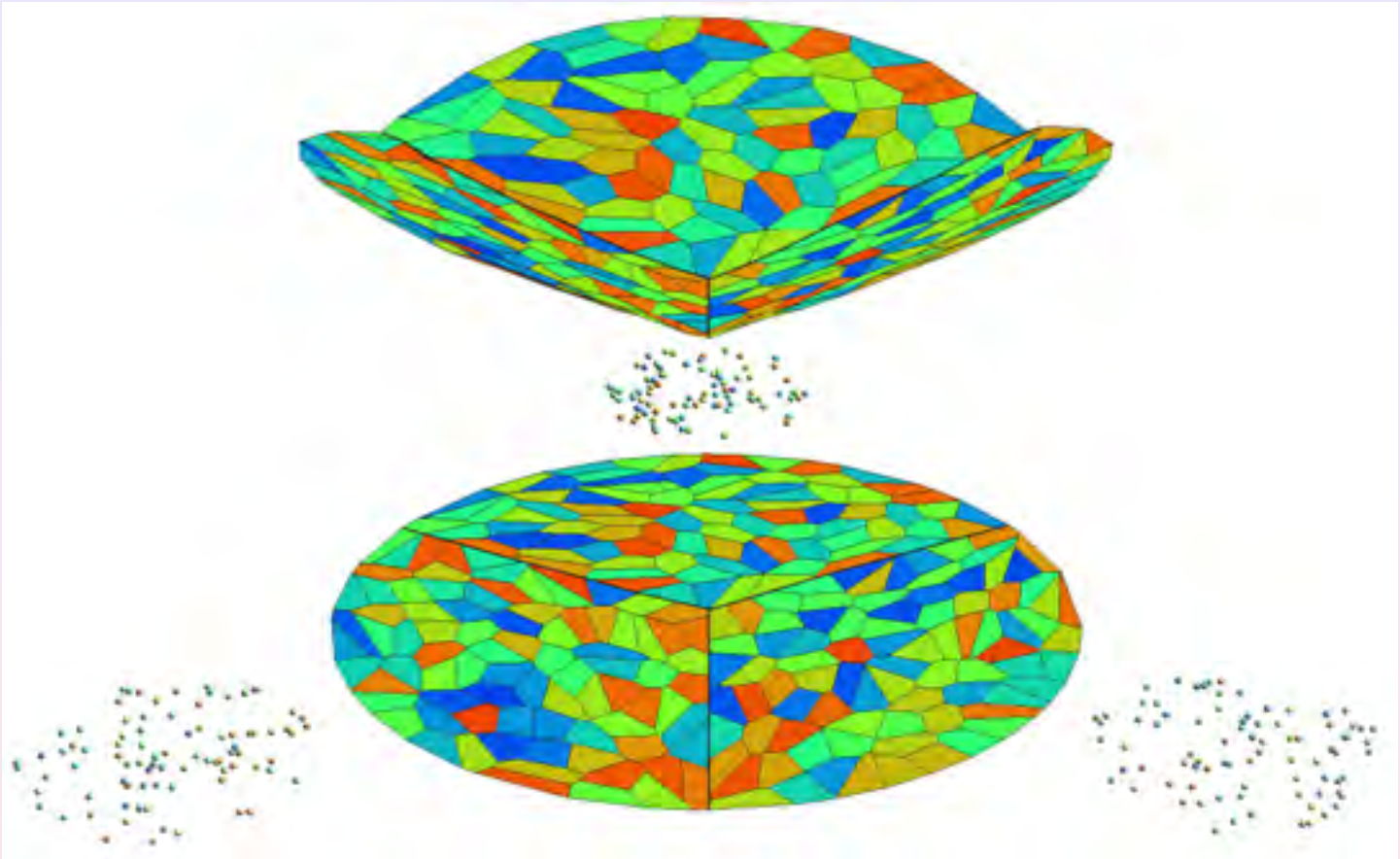


Figure: Optimal transportation map is discontinuous, but the Brenier potential itself is continuous. The projection of ridges are the discontinuity singular sets.

Discontinuity of Optimal Transportation Map

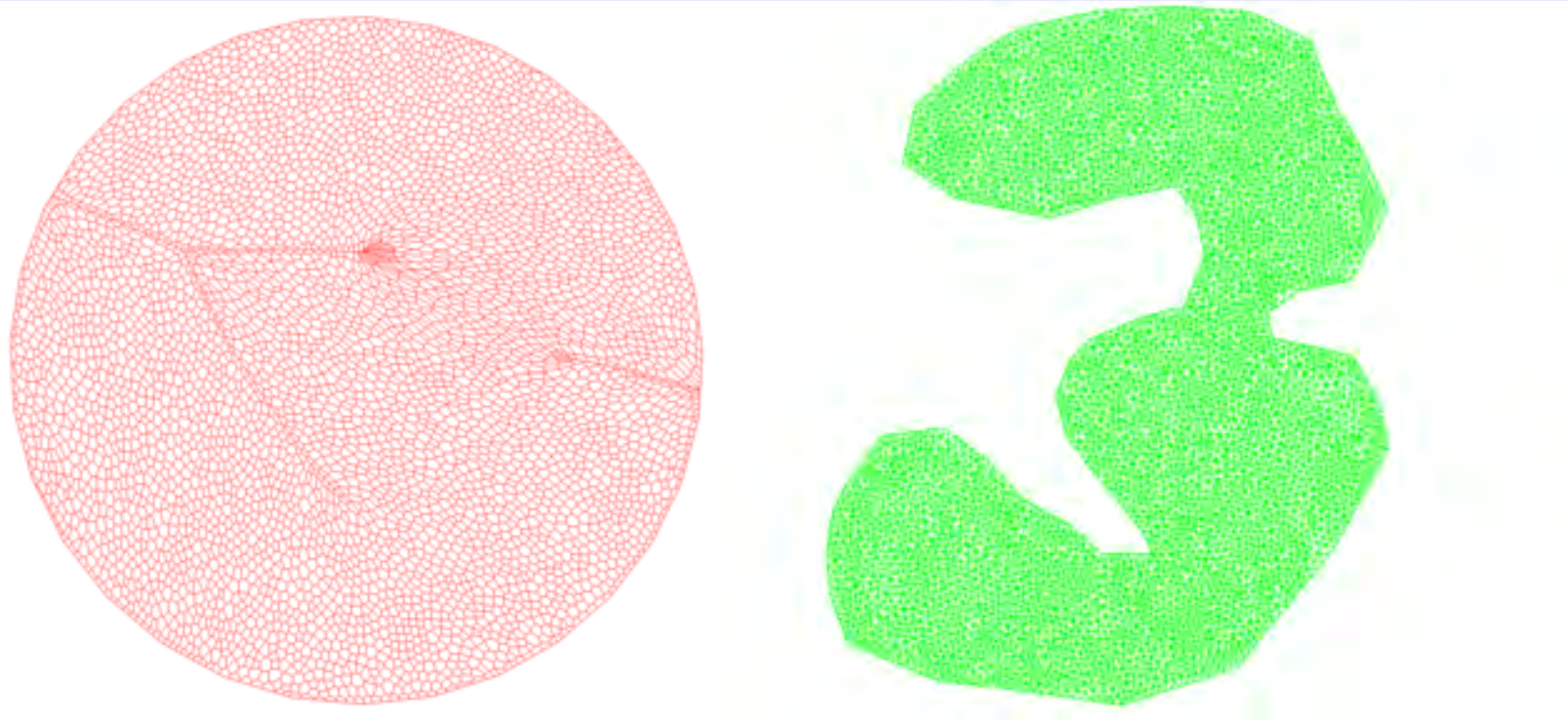


Figure: Optimal transportation map is discontinuous.

Discontinuity of Optimal Transportation Map

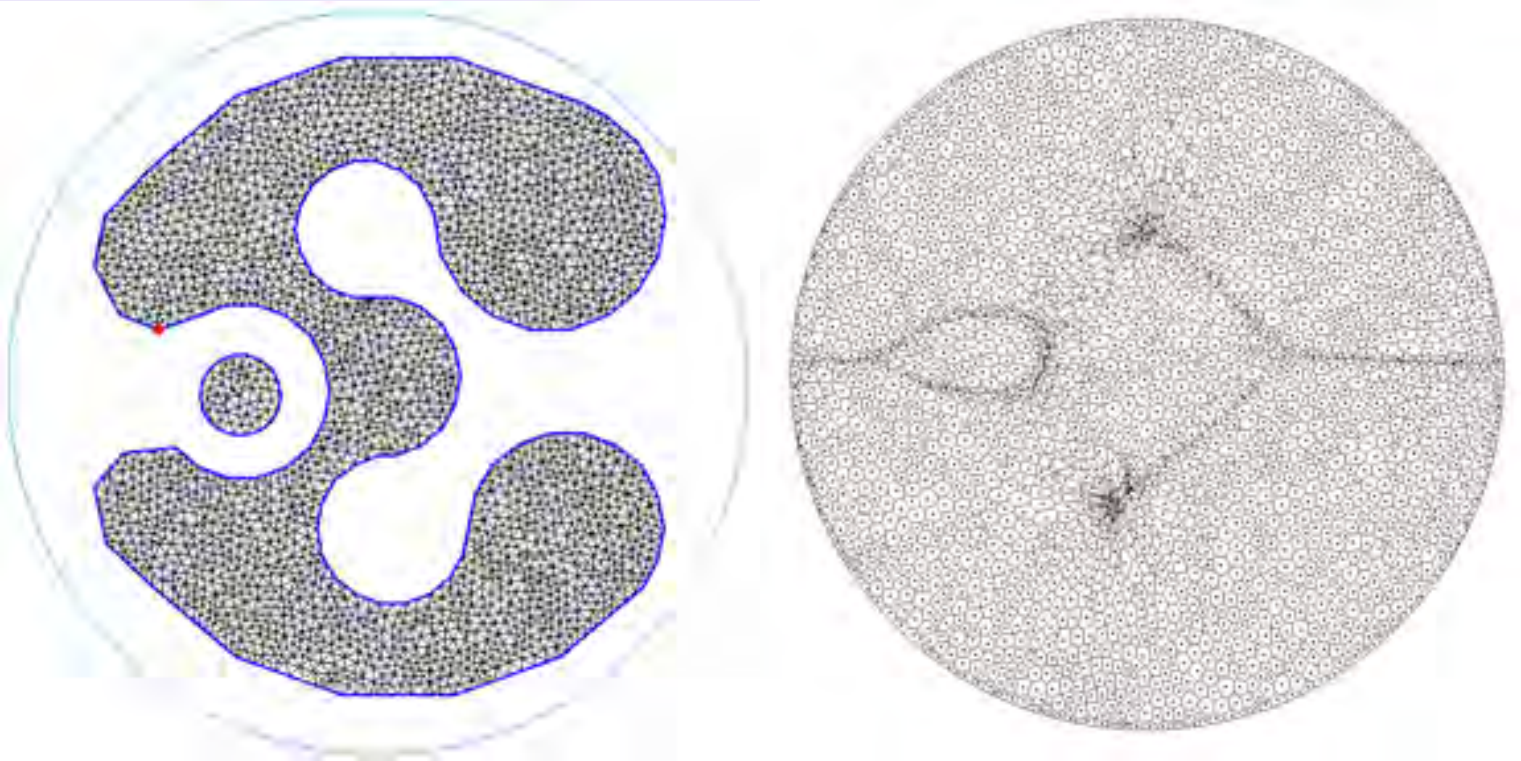


Figure: Optimal transportation map is discontinuous.

Mode Collapse

Mode Collapse

- GANs are difficult to train and sensitive to hyper-parameters;
- GANs suffer from mode collapsing, the generated distributions miss some modes;
- GANs may generate unrealistic samples;

Mode Collapse

1. The training process is unstable, and doesn't converge;
2. The searching converges to one of the multiple connected components of Λ, the mapping converges to one continuous branch of the desired transformation mapping. This means we encounter a mode collapse;
3. The training process leads to a transportation map, which covers all the modes successfully, but also cover the regions outside Λ. In practice, this will induce the phenomena of generating unrealistic samples.

Mode Collapse

Intrinsic Conflict

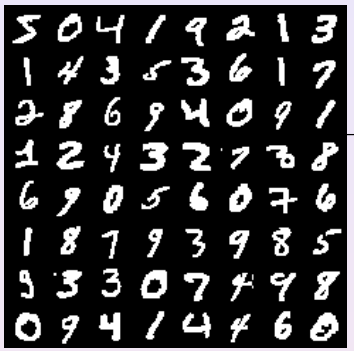
Deep neural networks can only represent continuous mappings, but the transportation maps are discontinuous on singular sets. Namely, the target mappings are outside the functional space of Dnns. This conflict induces mode collapsing.

Avoid Mode Collapse

The optimal transport map is discontinuous, but Brenier potential itself is continuous. The neural network should represent the Brenier potential, instead of its gradient, namely the transportation map.

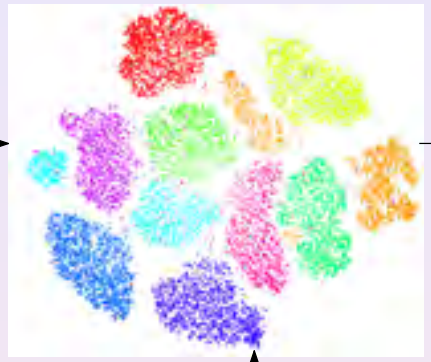
Avoid Mode Collapsing

Training Data



encoder

Latent Distribution



decoder

Generated Samples

OT Mapper

Brenier Potential

white noise

Distribution Transformation

Optimal Transportation Map

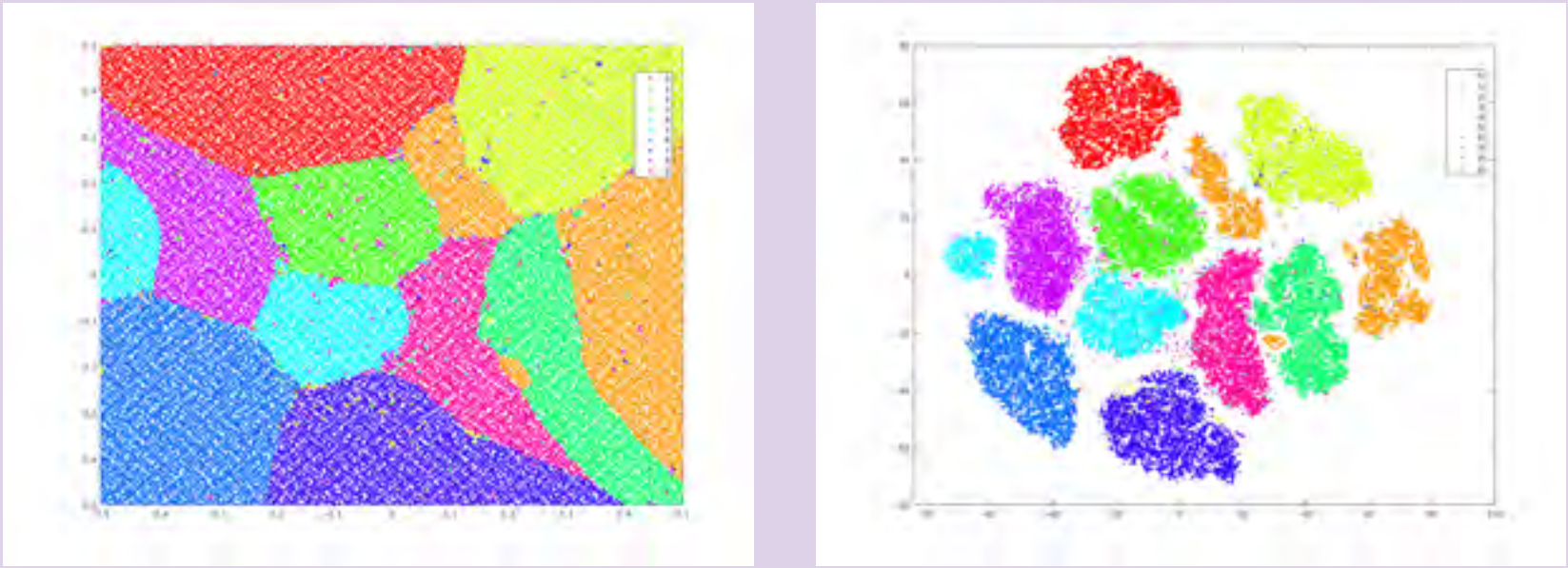


Figure: Find a mapping, which transform a source distribution to the data latent code distribution.

Optimal Transportation Map

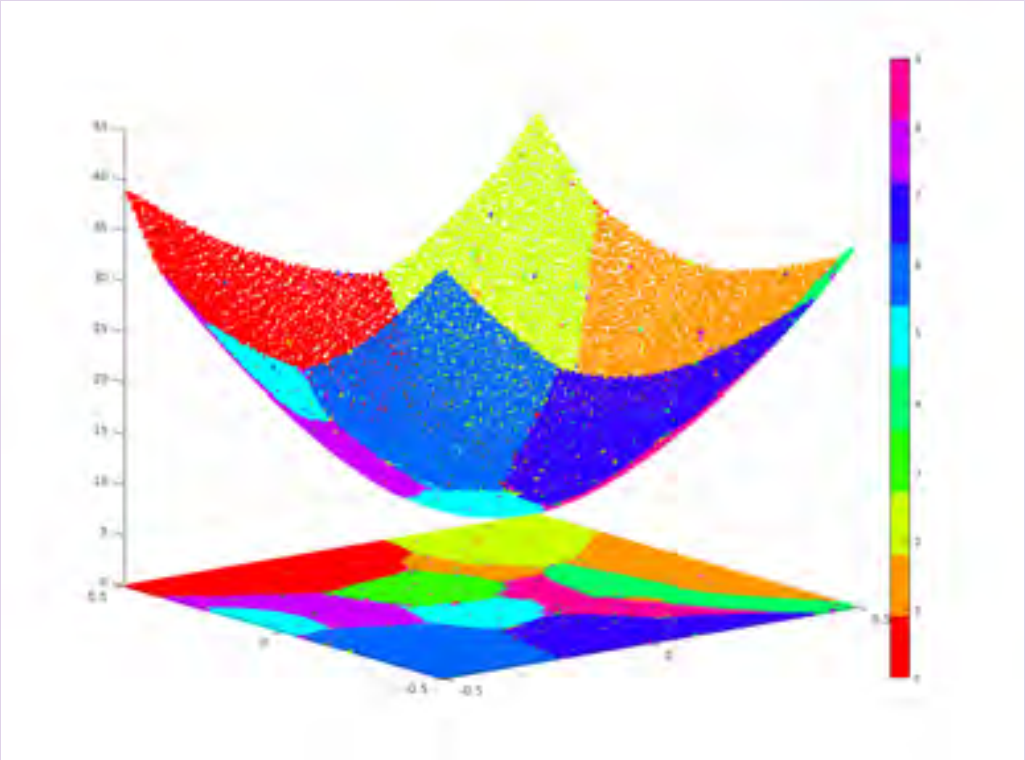


Figure: The optimal transportation map is given by a convex function u, $T = \nabla u$.

Singularity Set Detection

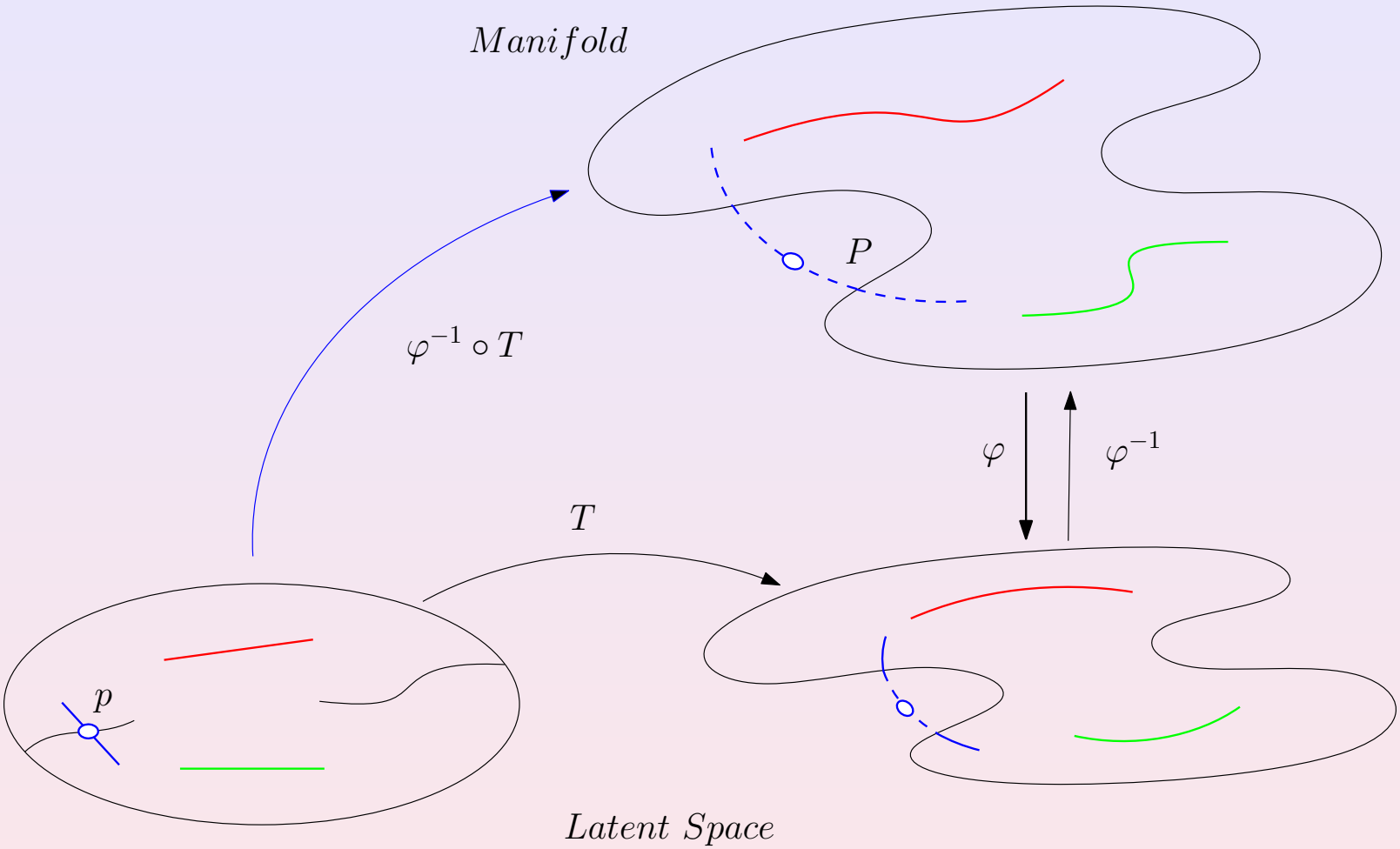


Figure: Singularity set detection.

Curves on facial photo manifold

Figure: Curves on facial photo manifold.

Mode Collapse

(a) generated facial images

(b) a path through a singularity.

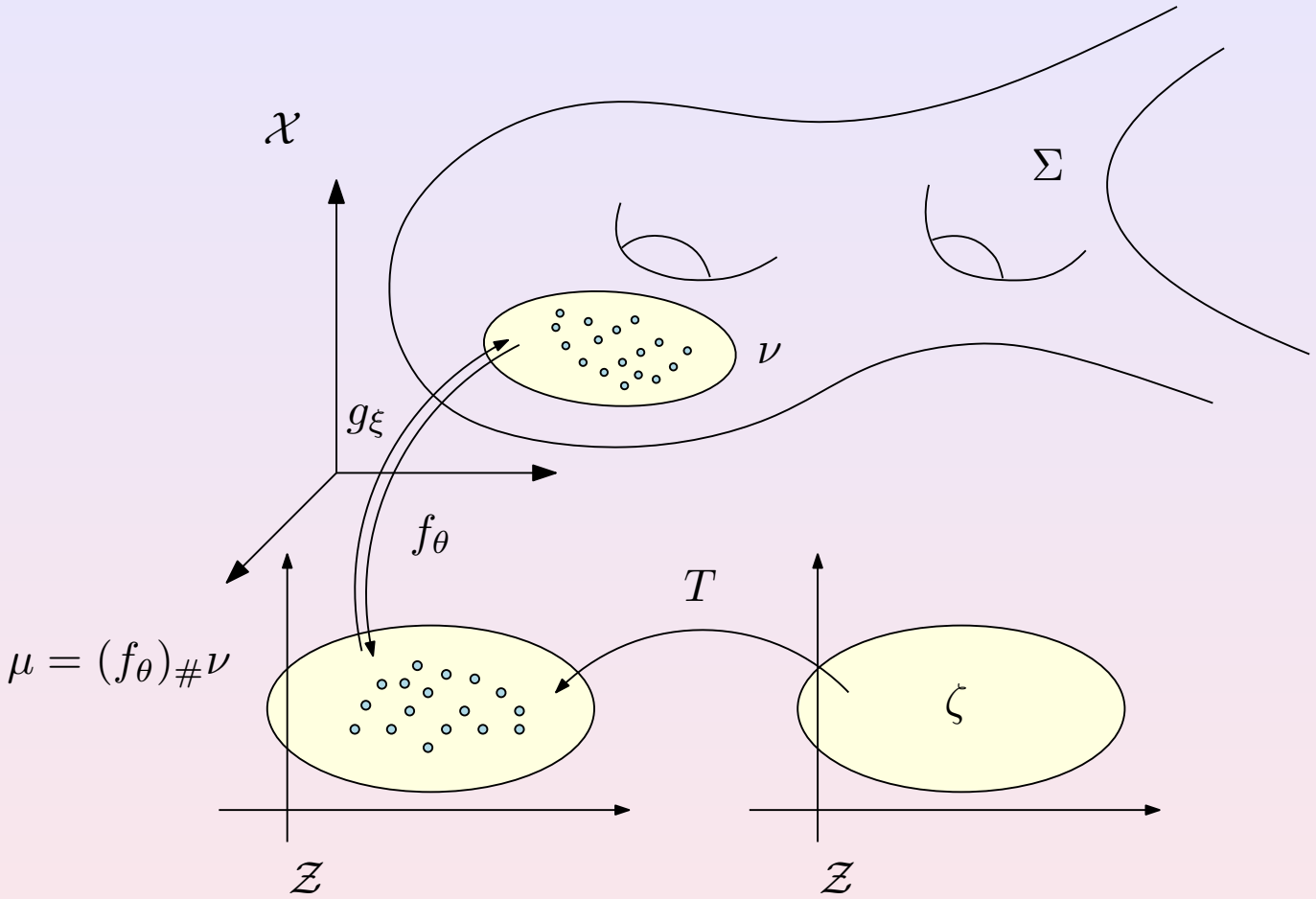
Figure: Facial images generated by an AE-OT model, the image in the center of (b) shows the transportation map is discontinuous.

Mode Collapse

Figure: Facial images with zero probability (From Internet).

Autoencoder-Optimal Transportation Framework

Auntoencoder-OMT



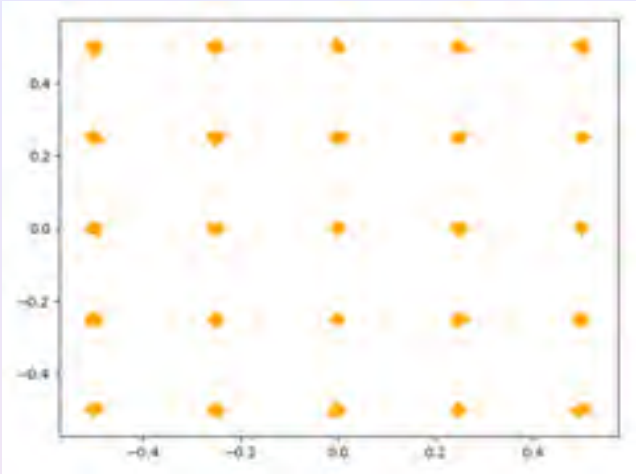
Use autoencoder to realize encoder and decoder, use OMT in the latent space to realize probability transformation.

Auntoencoder-OMT

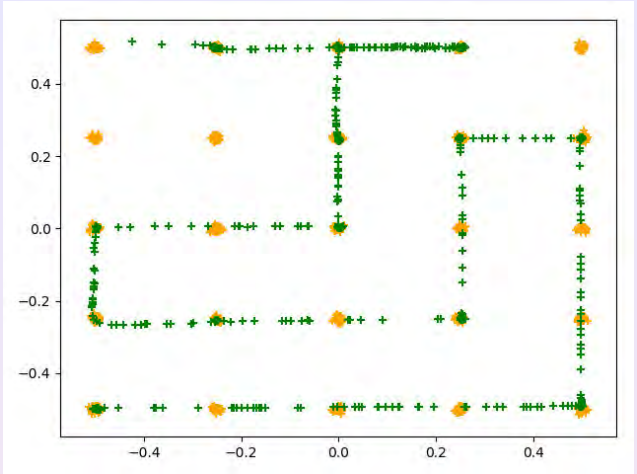
Merits

1. Solving Monge-Ampère equation is reduced to a convex optimization, which has unique solution. The optimization won't be trapped in a local optimum;
2. The Hessian matrix of the energy has explicit formulation. The Newton's method can be applied with second order convergence; or the quasi-Newton's method can be used with super-linear convergence. Whereas conventional gradient descend method has linear convergence;
3. The approximation accuracy can be fully controlled by the density of the sampling density by using Monte-Carlo method;
4. The algorithm can be refined to be hierarchical and self-adaptive to further improve the efficiency;
5. The parallel algorithm can be implemented using GPU.

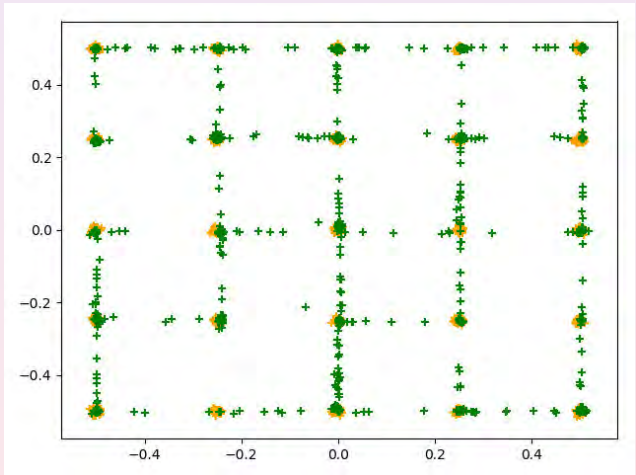
Experiments - Mode Collapse



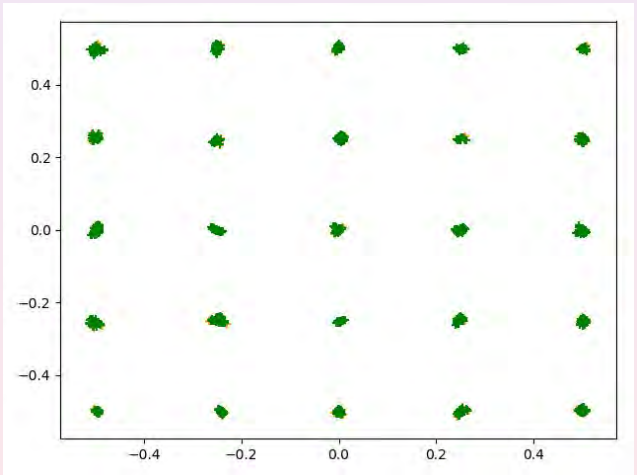
(a) original



(b) GAN



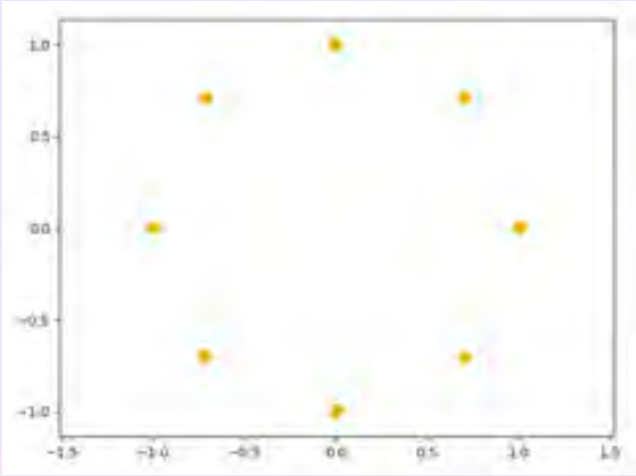
(c) pacgan



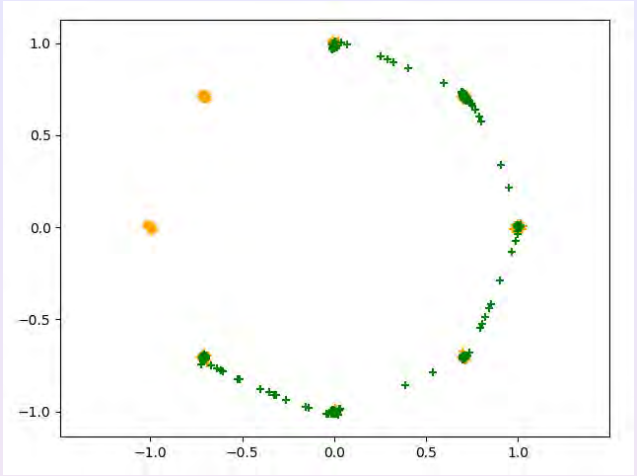
(d) Our model, AE-OT

Figure: Comparison between conventional models with AE-OT.

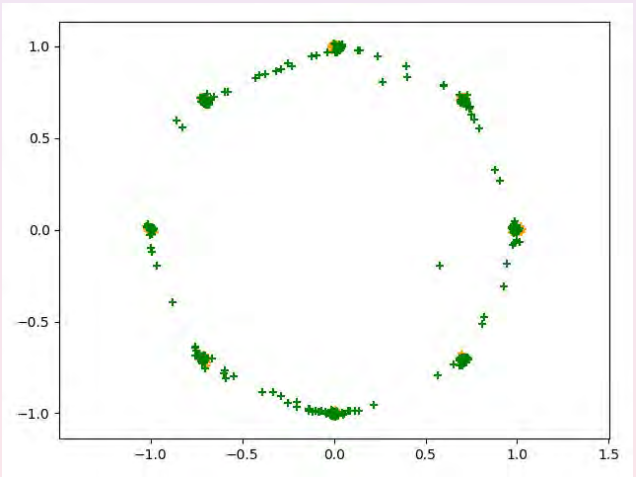
Experiments - Mode Collapse



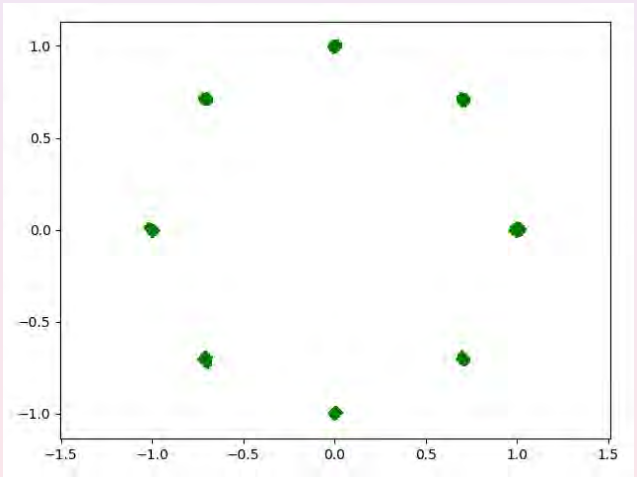
(a) original



(b) GAN



(c) pacgan



(d) Our model, AE-OT

Figure: Comparison between conventional models with AE-OT.

Experiments - mnist

(a) VAE

(b) WGAN

(c) Our model, AE-OT

(d) Our model, AE-OT

Figure: Comparison between conventional models VAE and WGAN with our model AE-OT (AutoEncoder-OptimalTransportation).

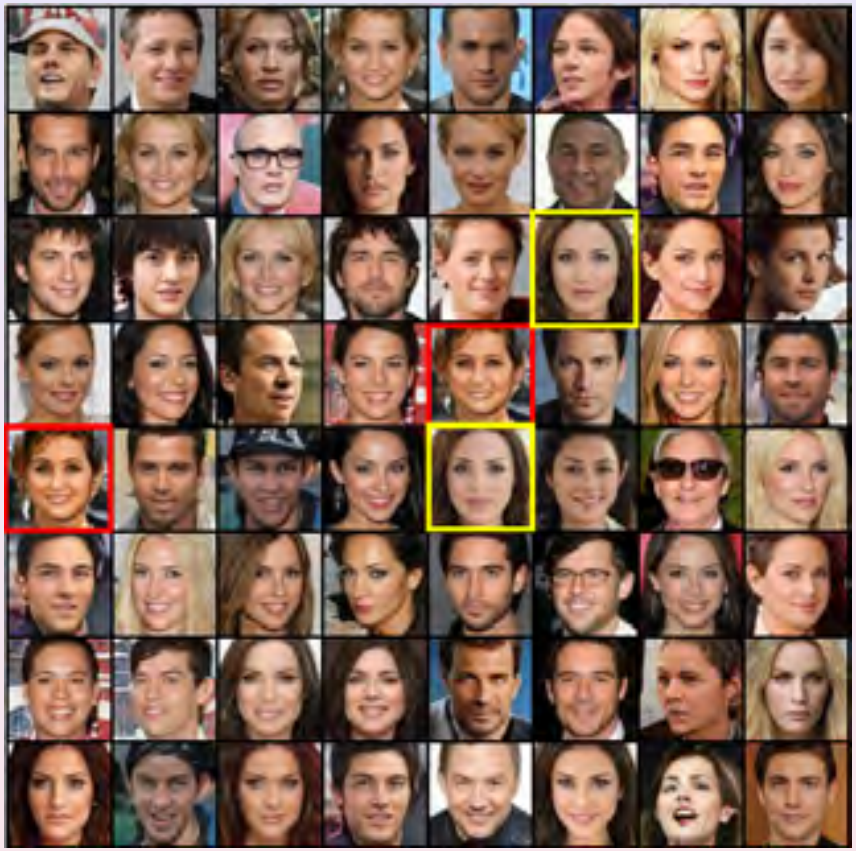
Experiments - WGAN-QC CelebA

(a) WGAN-GP

(b) WGAN-div

Figure: Failure cases for WGAN-GP and WGAN-div.

Experiments - WGAN-QC CelebA



(c) CRGAN - mode collapsing

(d) Our model

Figure: Comparision between CRGAN and our model.

Experiments - WGAN-QC CelebAHQ

Figure: Human facial images generated by our model.

Experiments - WGAN-QC CelebAHQ

Figure: Human facial images generated by our model.

Experiments - AE-OT Interpolation

Figure: Human facial images generated by our AE-OT model (AutoEncoder-OptimalTransportation).

Experiments - WGAN-QC Interpolation

Figure: Human facial images generated by our model.

Experiments - MNIST Fashion

MM GAN	NSGAN	LSGAN
29.6	26.5	30.7
WGAN	BEGAN	VAE
21.5	22.9	68.7
GLO	GLANN	AE-OMT
57.7	13	11.2

Figure: Our method has smallest FID score. (Fréchet Inception Distance)

Conclusion

This work introduces a geometric understanding of deep learning:

- The intrinsic pattern of natural data can be represented by manifold distribution principle.
- The deep learning system has two major tasks: manifold learning and probability distribution transformation.
- Optimal transportation assign a Riemannian metric in the space of distributions, so variational optimization can be carried out.
- By Brenier theory, the generator and discriminator should collaborate instead of compete with each other;
- The regularity theory of Monge-Ampere equation explains mode collapse;
- The AE-OT framework can avoid mode collapse, and make half the blackbox transparent.

References

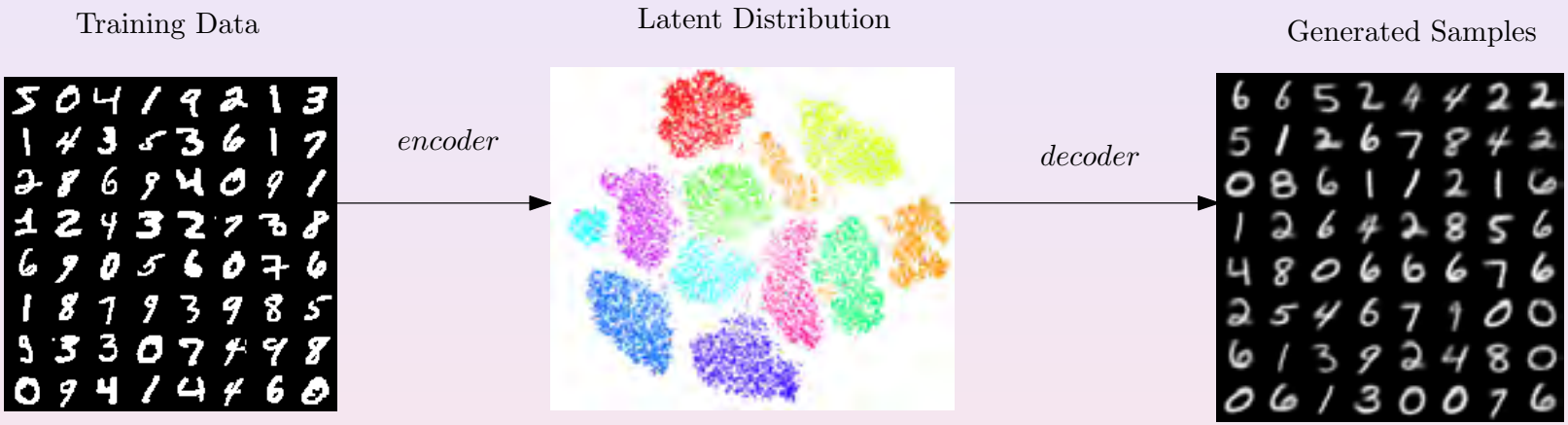
- X. Gu, F. Luo, J. Sun and S.-T. Yau, “Variational Principles for Minkowski Type Problems, Discrete Optimal Transport, and Discrete Monge-Ampere Equations”, AJM Volume 20, Number 2, Pages 383-398, 2016.
- D. An, Y. Guo, M. Zhang, X. Qin, N. Lei, X. Gu, “AE-OT-GAN”, ECCV 2020.
- D. An, Y. Guo, N. Lei, Z. Luo, S.-T. Yau and X. Gu, “AE-OT: A New Generative Model Based on Extended Semi-Discrete Optimal Transport”, ICLR2020.
- H. Liu, X. Gu and D. Samaras, “Wasserstein GAN with Quadratic Transport Cost”, ICCV 2019.
- H. Liu, X. Gu and D. Samaras, “A Two-Step Computation of the Exact GAN Wasserstein Distance”, ICML 2018.
- N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau and X. Gu, “Geometric Understanding of Deep Learning”, Journal of Engineering, 2020.
- N. Lei, K. Su, L. Cui, S.-T. Yau and X. Gu, “A Geometric View of Optimal Transportation and Generative Model”, CAGD, 68(2019),1-21.

Thanks

Please email to gu@cmsa.math.harvard.edu, gu@cs.stonybrook.edu.

Thank you!

AutoEncoder Framework



Online Course

Please email to gu@cs.stonybrook.edu,gu@cmsa.fas.harvard.edu.

online course!

Subgradient

Definition (subgradient)

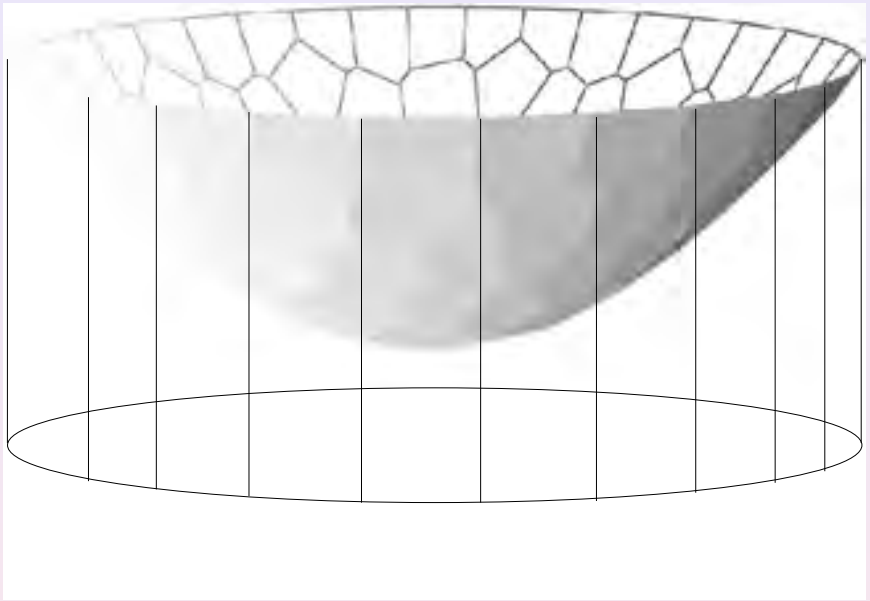
Given an open set $\Omega \subset \mathbb{R}^d$ and $u : \Omega \to \mathbb{R}$ a convex function, for $x \in \Omega$, the subgradient (subdifferential) of u at x is defined as

$$\partial u(x) := \{p \in \mathbb{R}^n : u(z) \geq u(x) + \langle p, z - x \rangle \quad \forall z \in \Omega\}.$$

The Brenier potential u is differentiable at x if its subgradient $\partial u(x)$ is a singleton. We classify the points according to the dimensions of their subgradients, and define the sets

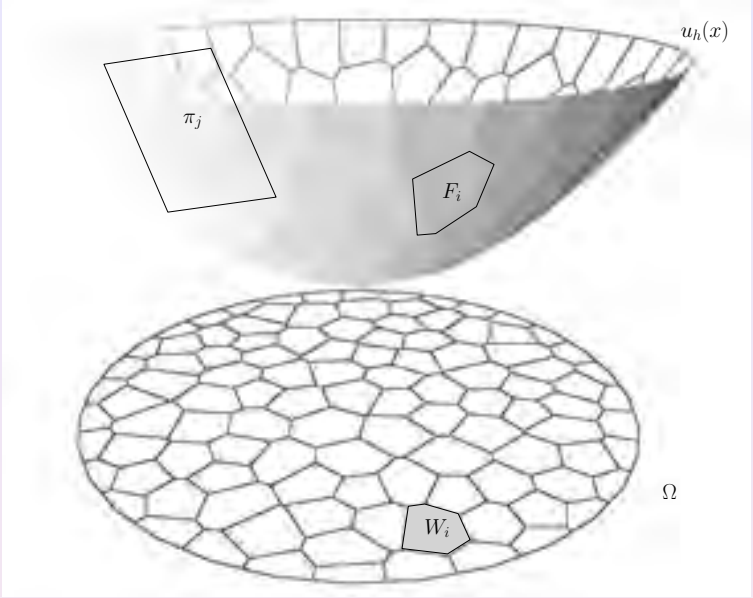
$$\Sigma_k(u) := \left\{ x \in \mathbb{R}^d | \ \dim(\partial u(x)) = k \right\}, \quad k = 0, 1, 2 \ldots, d.$$

Geometric Interpretation



One can define a cylinder through $\partial\Omega$, the cylinder is truncated by the xy-plane and the convex polyhedron. The energy term $\int^{\mathbf{h}} \sum w_i(\eta) d\eta_i$ equals to the volume of the truncated cylinder.

Computational Algorithm



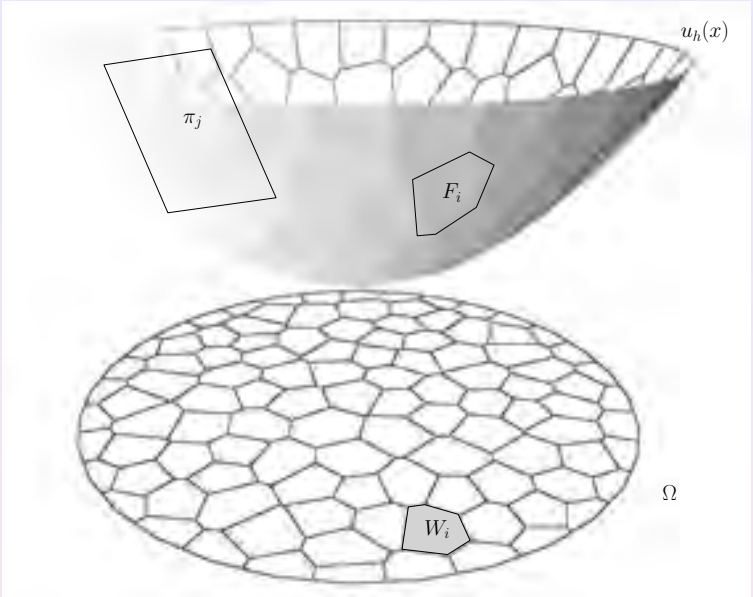
Definition (Alexandrov Potential)

The concave energy is

$$E(h_1, h_2, \cdots, h_k) = \sum_{i=1}^{k} v_i h_i - \int_{\mathbf{0}}^{\mathbf{h}} \sum_{j=1}^{k} w_j(\eta) d\eta_j,$$

Geometrically, the energy is the volume beneath the parabola.

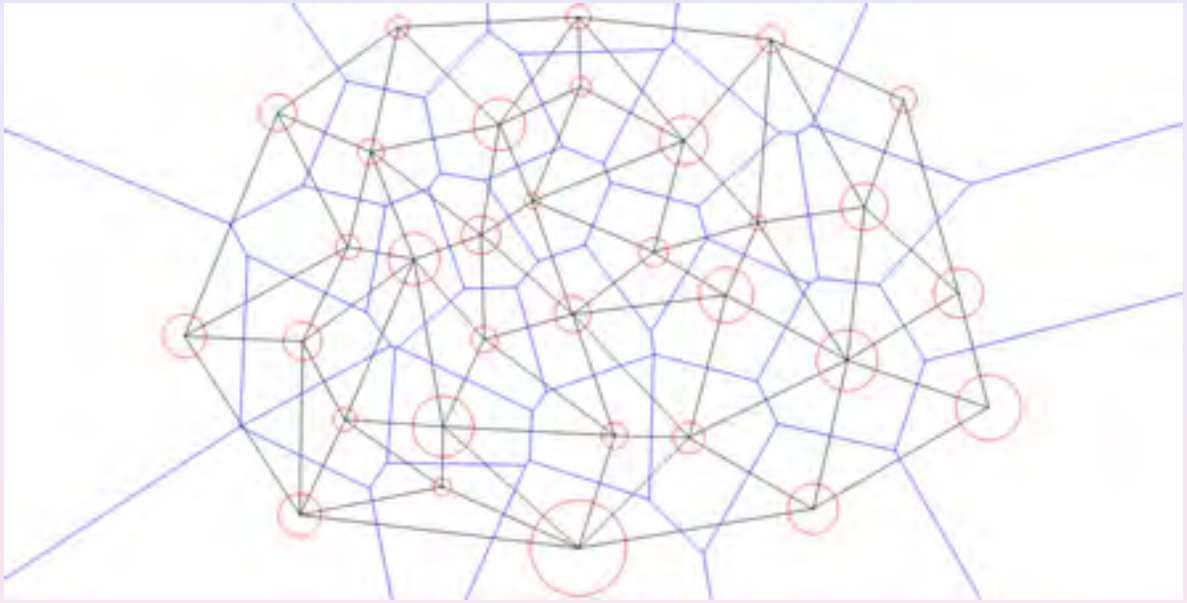
Computational Algorithm



The gradient of the Alexanrov potential is the differences between the target measure and the current measure of each cell

$$\nabla E(h_1, h_2, \cdots, h_k) = (\nu_1 - w_1, \nu_2 - w_2, \cdots, \nu_k - w_k)$$

Computational Algorithm



The Hessian of the energy is the length ratios of edge and dual edges,

$$\frac{\partial w_i}{\partial h_j} = \frac{|e_{ij}|}{|\bar{e}_{ij}|}$$

Experiments - AE-OT CelebA

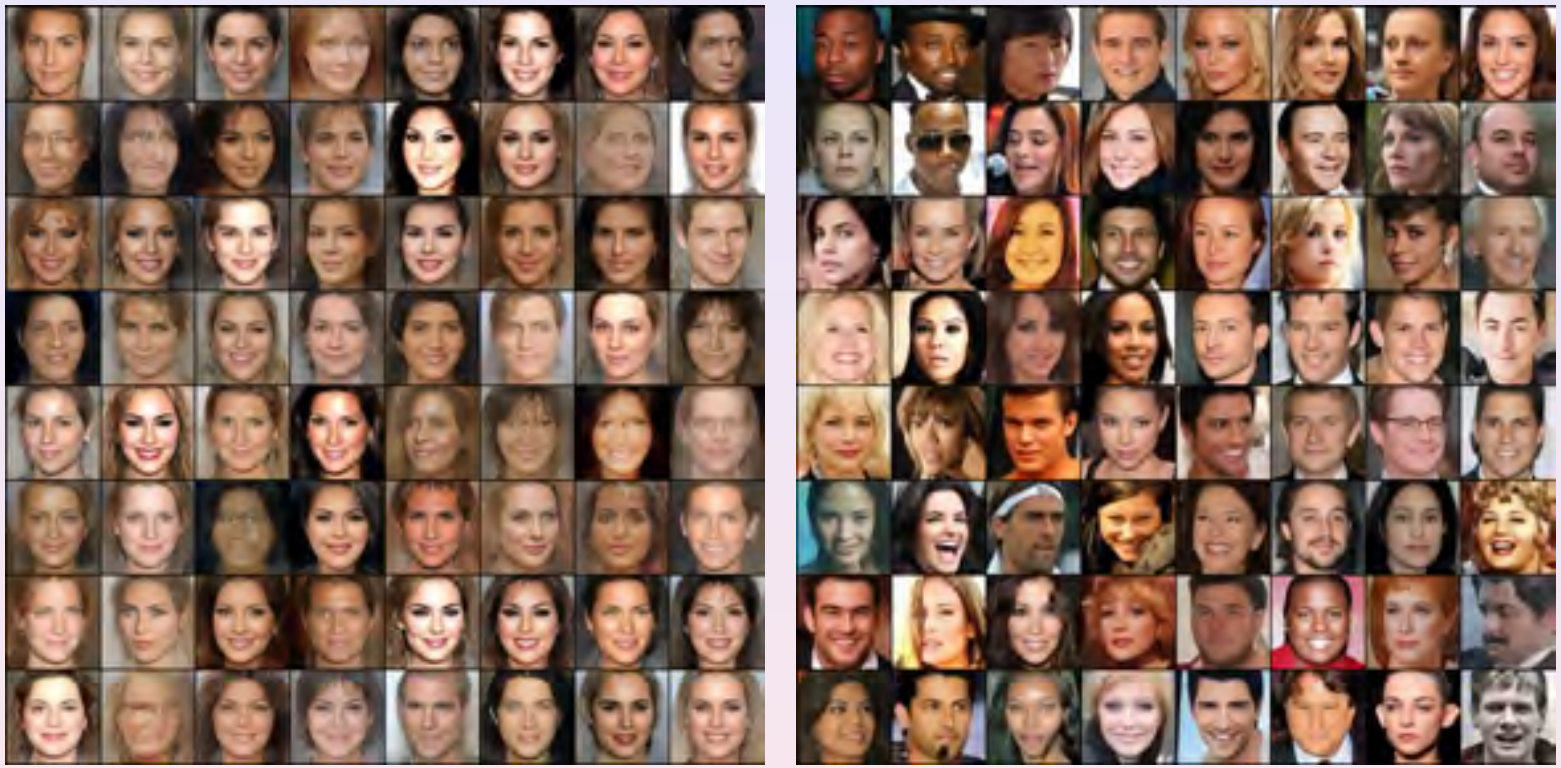


Figure: Human facial images generated by our AE-OT model (AutoEncoder-OptimalTransportation).