A Geometric View to Optimal Transportation and Generative Model

David Xianfeng Gu[1]

[1]Computer Science & Applied Mathematics
SUNY at Stony Brook University

Center of Mathematical Sciences and Appications
Harvard University

Geometric Computation and Applications
Trinity College, Dublin, Ireland

Thanks

Thanks for the invitation.

Collaborators

These projects are collaborated with Shing-Tung Yau, Feng Luo, Zhongxuan Luo, Na Lei, Dimitris Samaras and so on.

Outline

1. Why dose DL work?
2. How to quantify the learning capability of a DNN?
3. How does DL manipulate the probability distributions?

Why dose DL work?

Deep Learning

Deep learning is the mainstream technique for many machine learning tasks, including image recognition, machine translation, speech recognition, and so on. Despite its success, the theoretical understanding on how it works remains primitive.

Manifold Assumption

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

Manifold Assumption

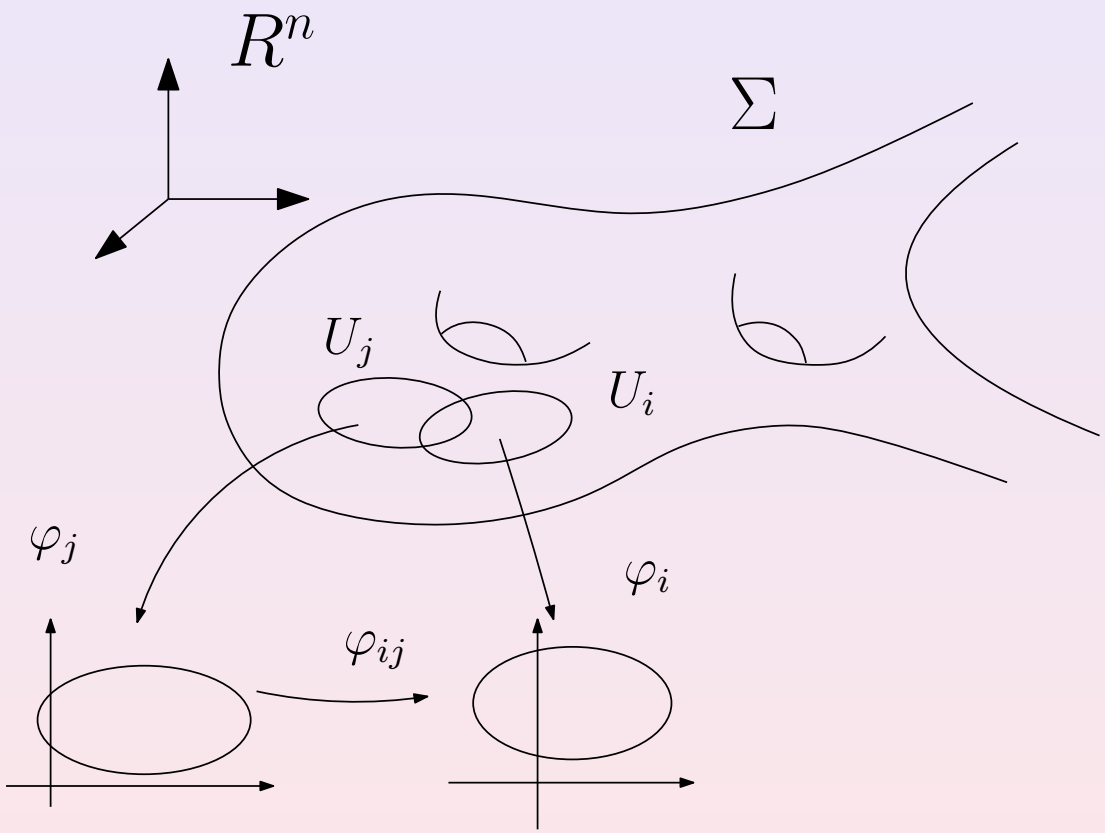
Natural high dimensional data concentrates close to a non-linear low-dimensional manifold.

Clustering Assumption

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

Deep learning method can learn and represent the manifold structure, and transform the probability distributions.

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

Manifold Structure

Definition (Manifold)

Suppose M is a topological space, covered by a set of open sets $M \subset \bigcup_\alpha U_\alpha$. For each open set U_α, there is a homeomorphism $\varphi_\alpha : U_\alpha \to \mathbb{R}^n$, the pair $(U_\alpha, \varphi_\alpha)$ form a chart. The union of charts form an atlas $\mathscr{A} = \{(U_\alpha, \varphi_\alpha)\}$. If $U_\alpha \cap U_\beta \neq \emptyset$, then the chart transition map is given by

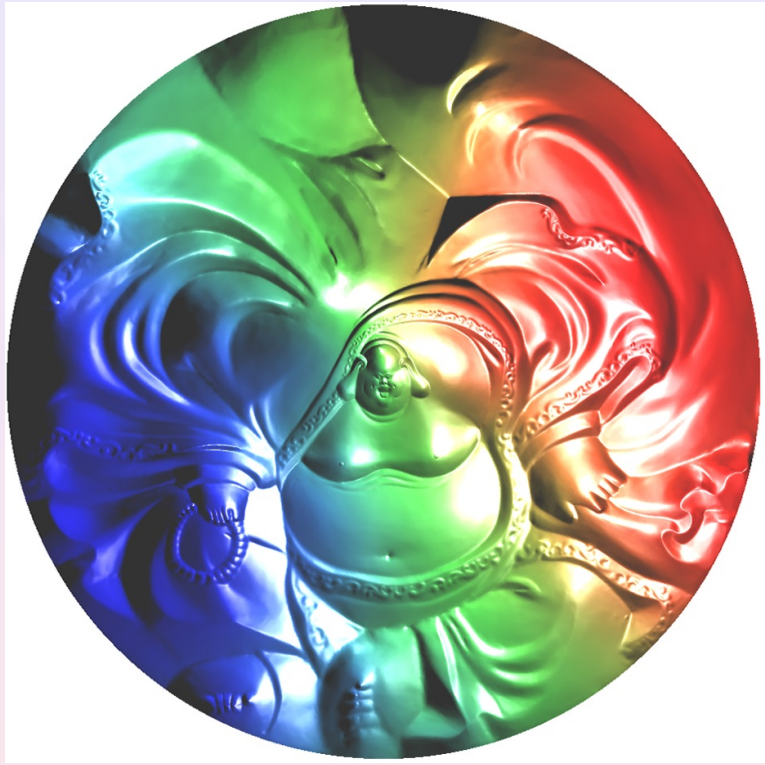
$$\varphi_{\alpha\beta} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta),$$

$$\varphi_{\alpha\beta} := \varphi_\beta \circ \varphi_\alpha^{-1}.$$

Example

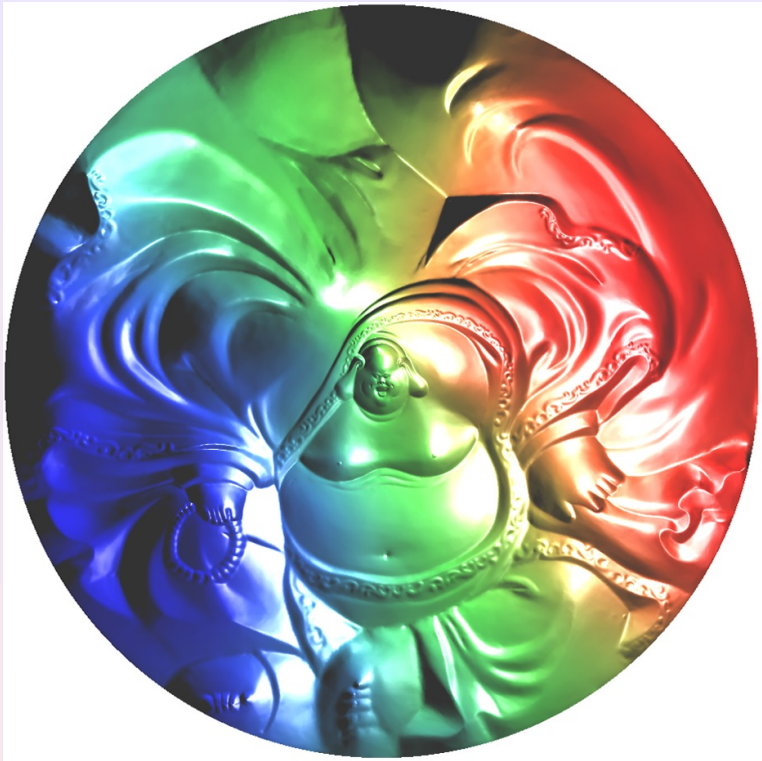
Image space $\mathscr{X}$ is $\mathbb{R}^3$; the data manifold Σ is the happy buddaha.

Example



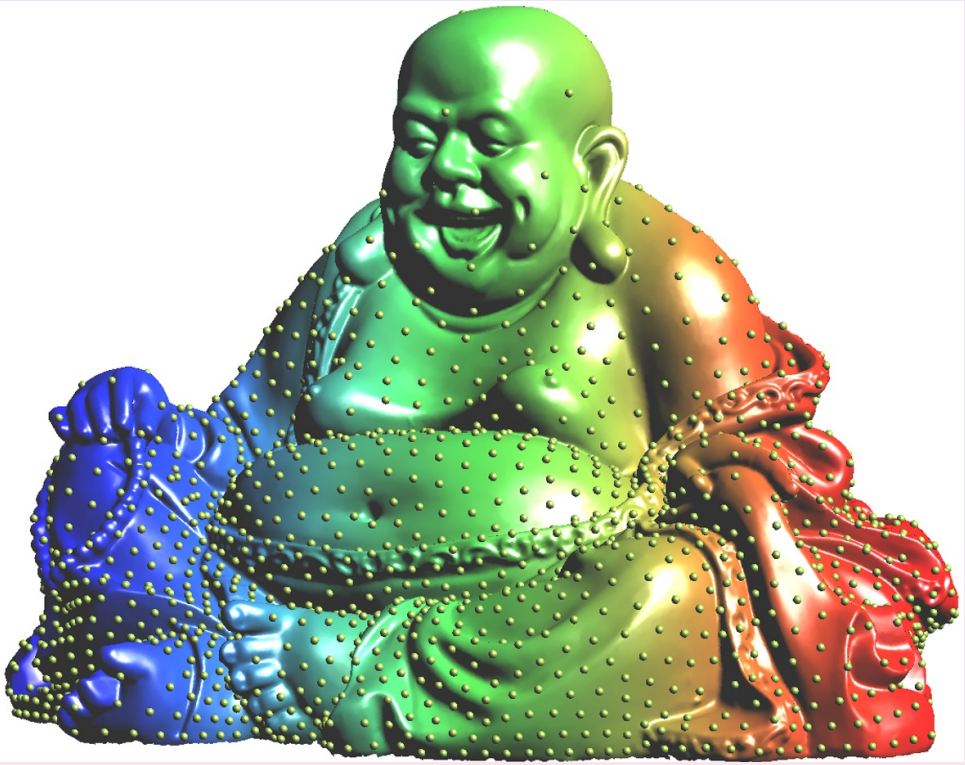
The encoding map is $\varphi_i : \Sigma \to \mathscr{Z}$; the decoding map is $\varphi_i^{-1} : \mathscr{Z} \to \Sigma$.

Example



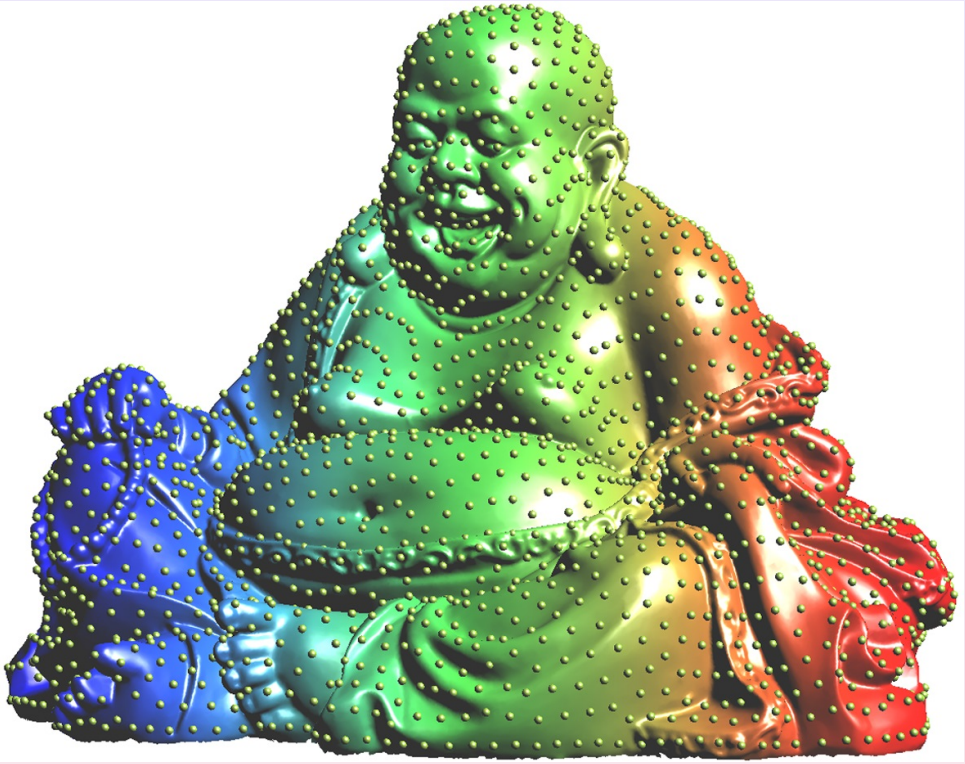
The automorphism of the latent space $\varphi_{ij} : \mathscr{Z} \to \mathscr{Z}$ is the chart transition.

Example



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

Example

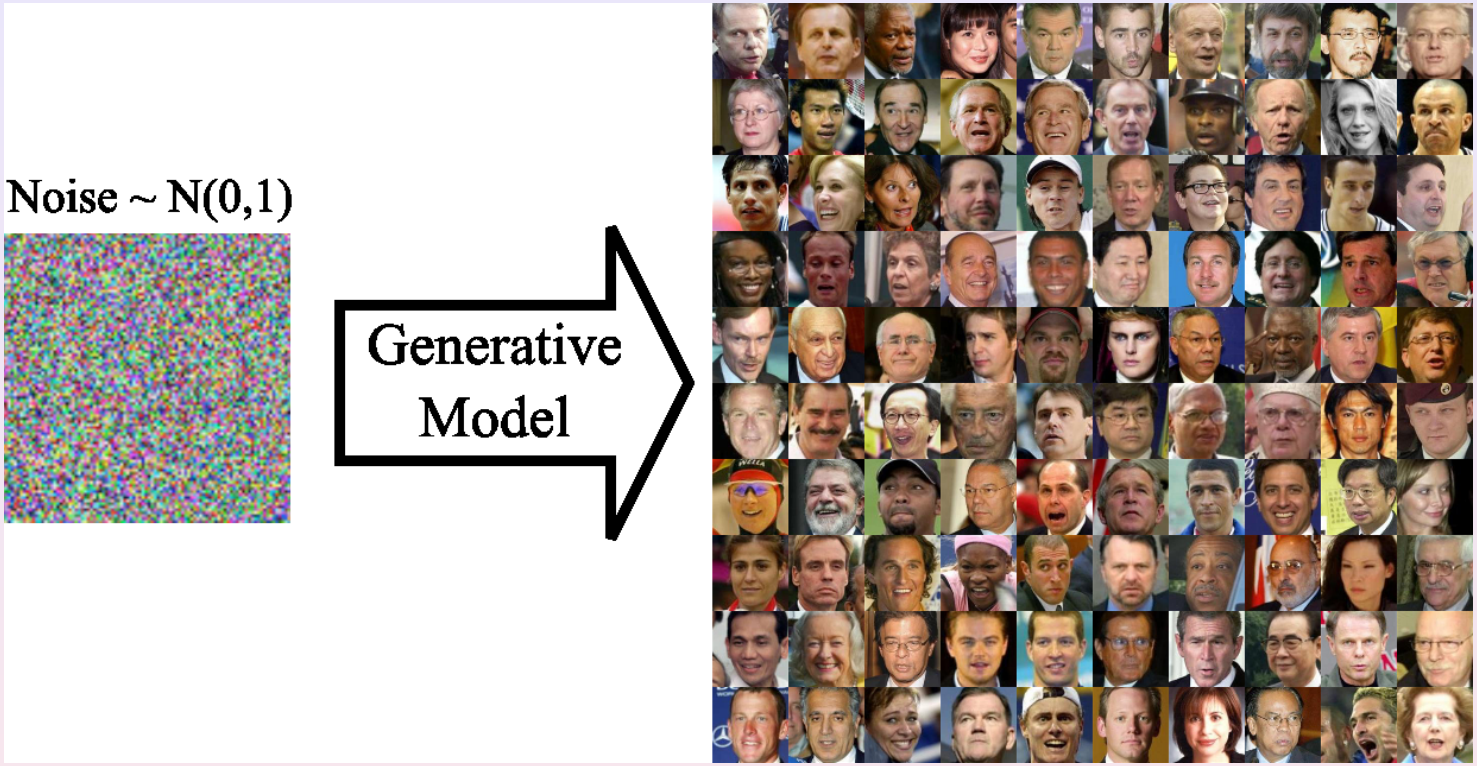


Uniform distribution ζ on the latent space $\mathscr{Z}$, uniform distribution on Σ produced by another decoding map.

Human Facial Image Manifold

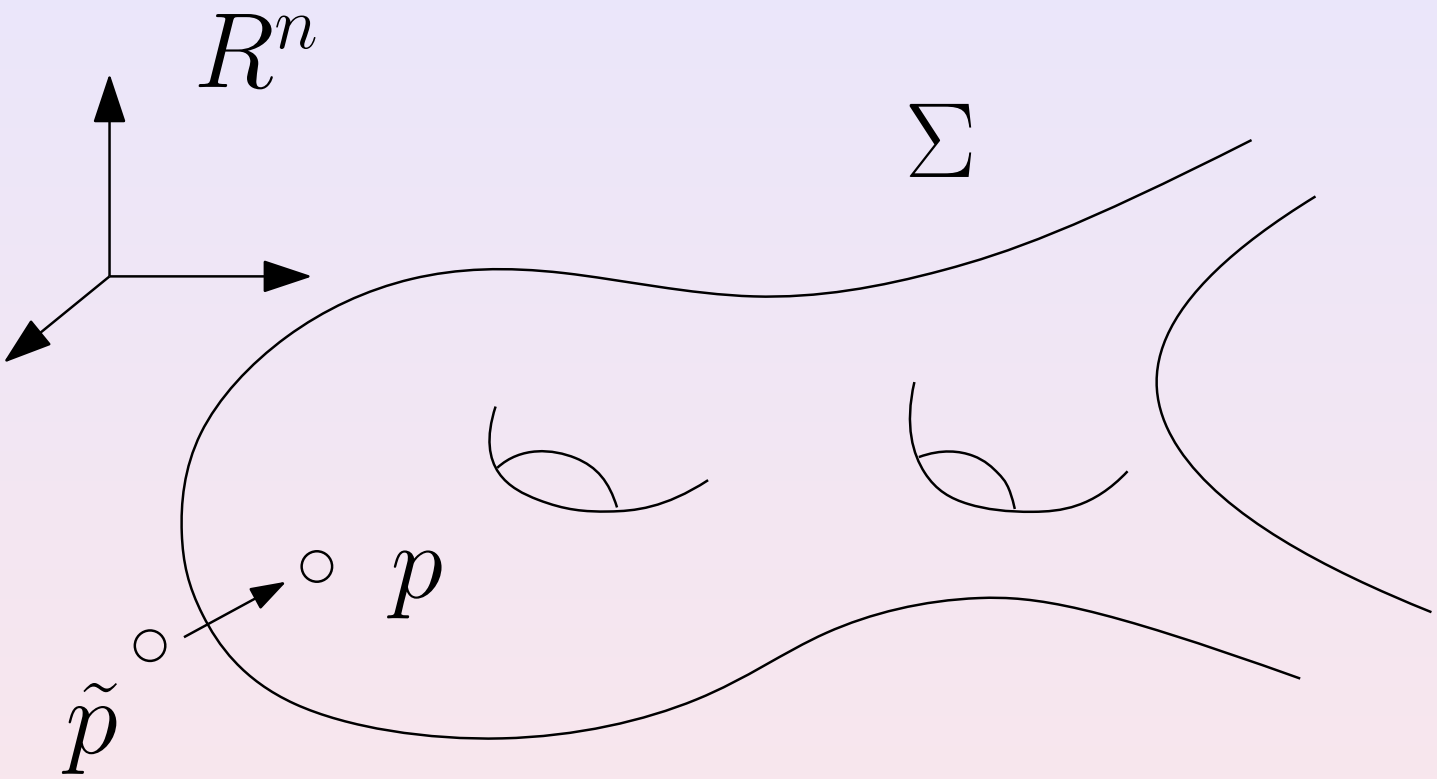
One facial image is determined by a finite number of genes, lighting conditions, camera parameters, therefore all facial images form a manifold.

Manifold view of Generative Model



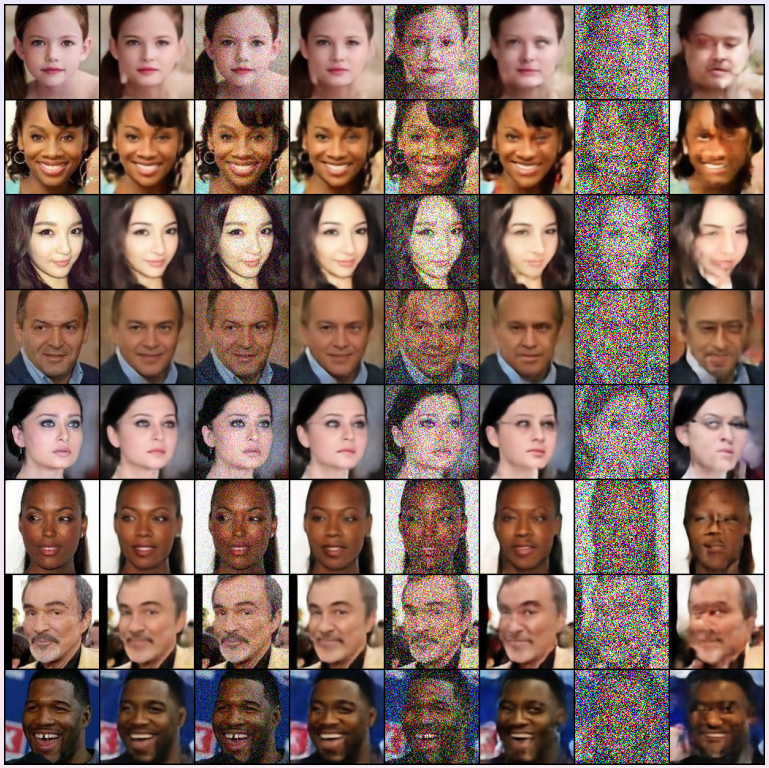
Given a parametric representation $\varphi : \mathscr{Z} \to \Sigma$, randomly generate a parameter $z \in \mathscr{Z}$ (white noise), $\varphi(z) \in \Sigma$ is a human facial image.

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.

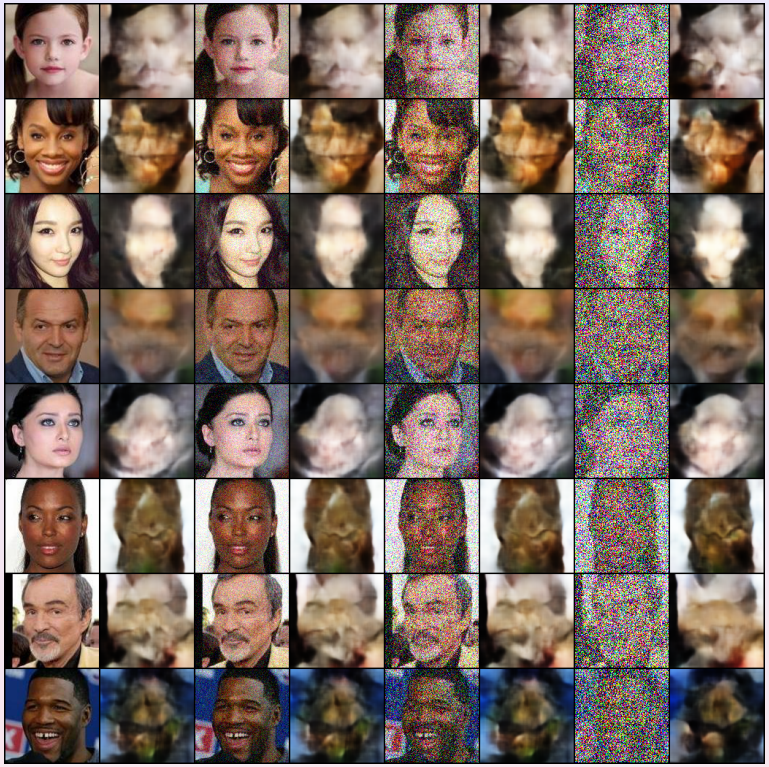
ML Method

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.

How dose DL learn a manifold?

Learning Task

The central tasks for Deep Learning are

1. Learn the manifold structure from the data;
2. Represent the manifold implicitly or explicitly.

Autoencoder

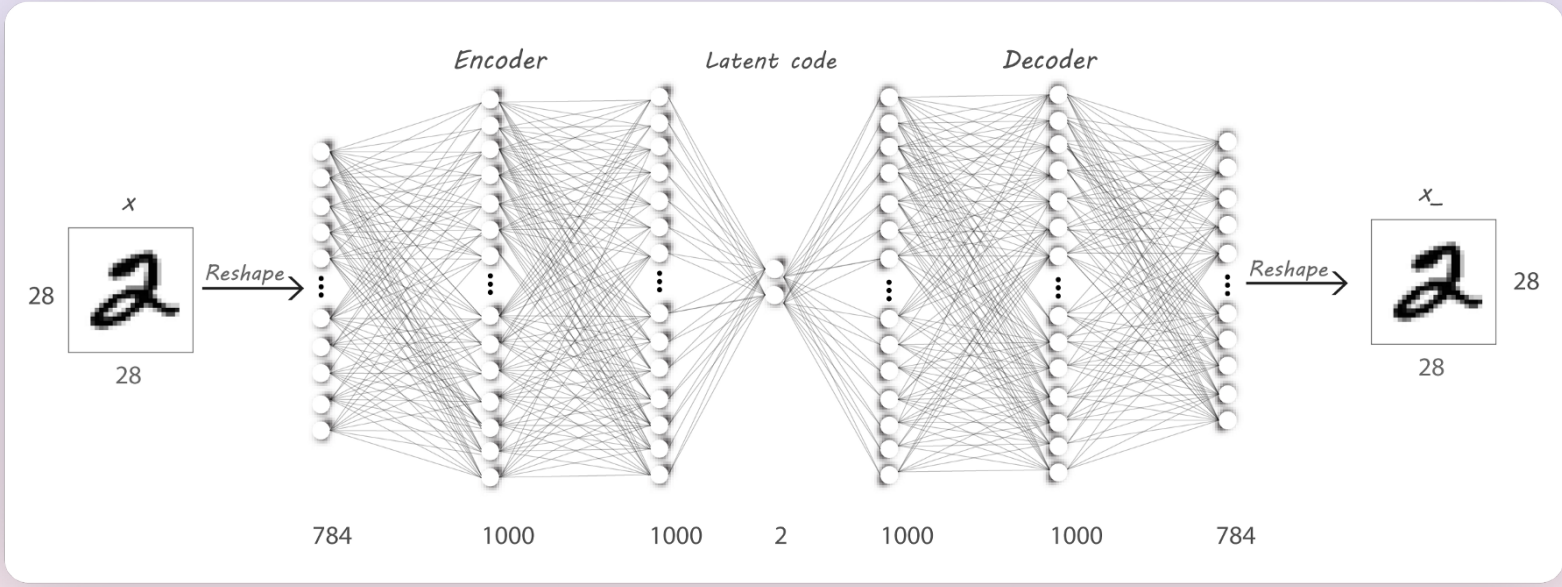
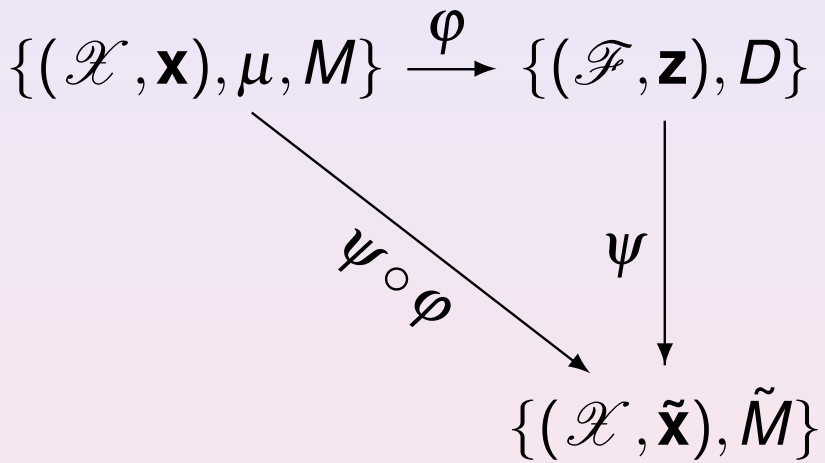


Figure: Auto-encoder architecture.

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

Autoencoder

The encoder takes a sample $\mathbf{x} \in \mathscr{X}$ and maps it to $\mathbf{z} \in \mathscr{F}$, $\mathbf{z} = \varphi(\mathbf{x})$. The decoder $\psi : \mathscr{F} \to \mathscr{X}$ maps $\mathbf{z}$ to the reconstruction $\tilde{\mathbf{x}}$.



An autoencoder is trained to minimise reconstruction errors:

$$\varphi, \psi = \text{argmin}_{\varphi,\psi} \int_{\mathscr{X}} \mathscr{L}(\mathbf{x}, \psi \circ \varphi(\mathbf{x})) d\mu(\mathbf{x}),$$

where $\mathscr{L}(\cdot, \cdot)$ is the loss function, such as squared errors. The reconstructed manifold $\tilde{M} = \psi \circ \varphi(M)$ is used as an approximation of M.

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 RNN, 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 $\mathbf{x}_1$ $\mathbf{x}_2$. $\mathscr{D}(\varphi_\theta)$ 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.

Encoding/Decoding

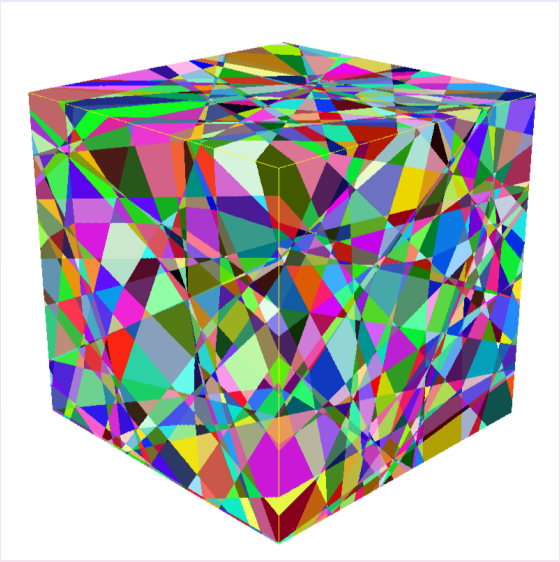
a. Input manifold
$M \subset \mathscr{X}$

b. latent representation
$D = \varphi_\theta(M)$

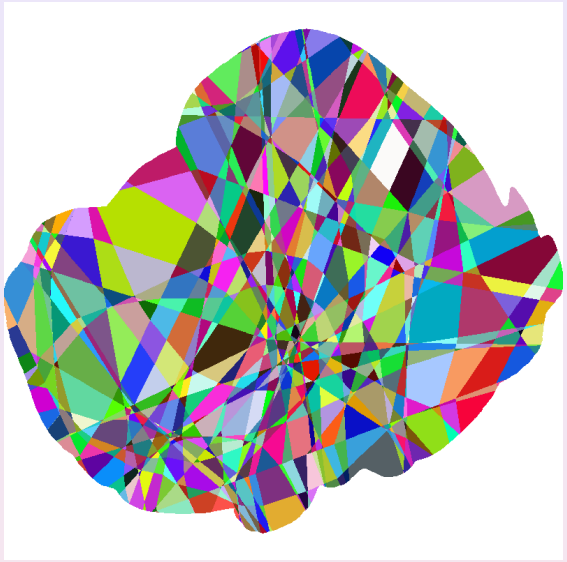
c. reconstructed mfld
$\tilde{M} = \psi_\theta(D)$

Figure: Auto-encoder pipeline.

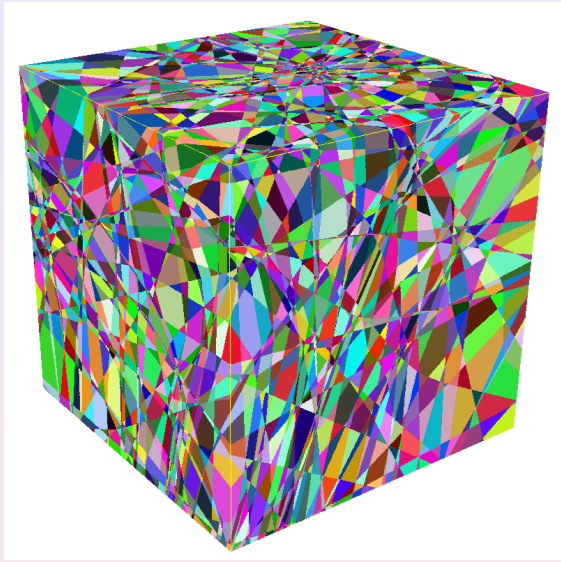
Piecewise Linear Mapping



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.

RL Complexity of a DNN

Definition (Rectified Linear Complexity of a ReLU DNN)

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,

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

Rectified Linear complexity 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

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

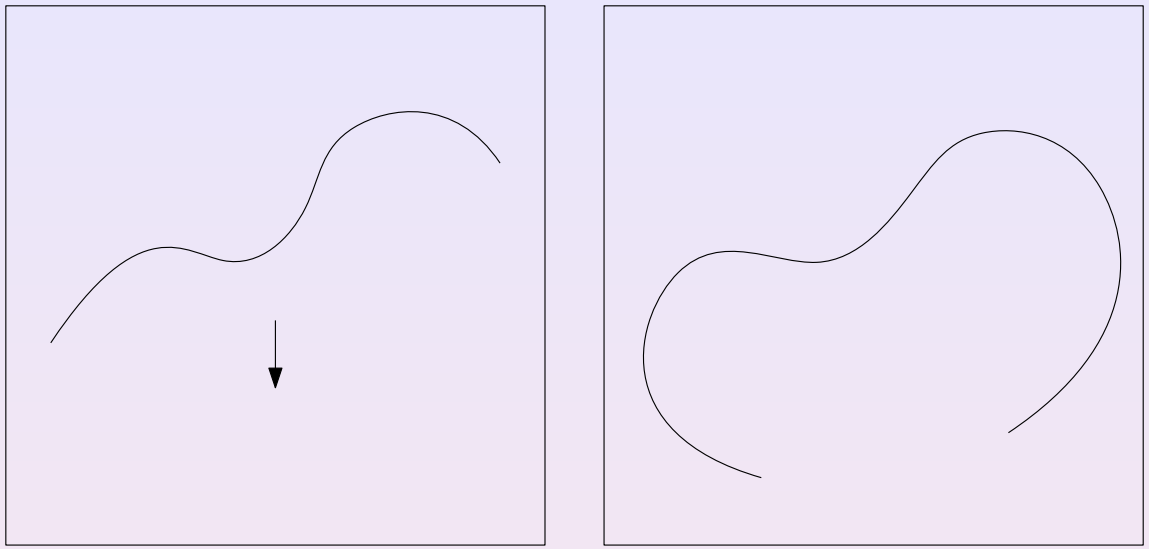
RL Complexity Upper Bound

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)

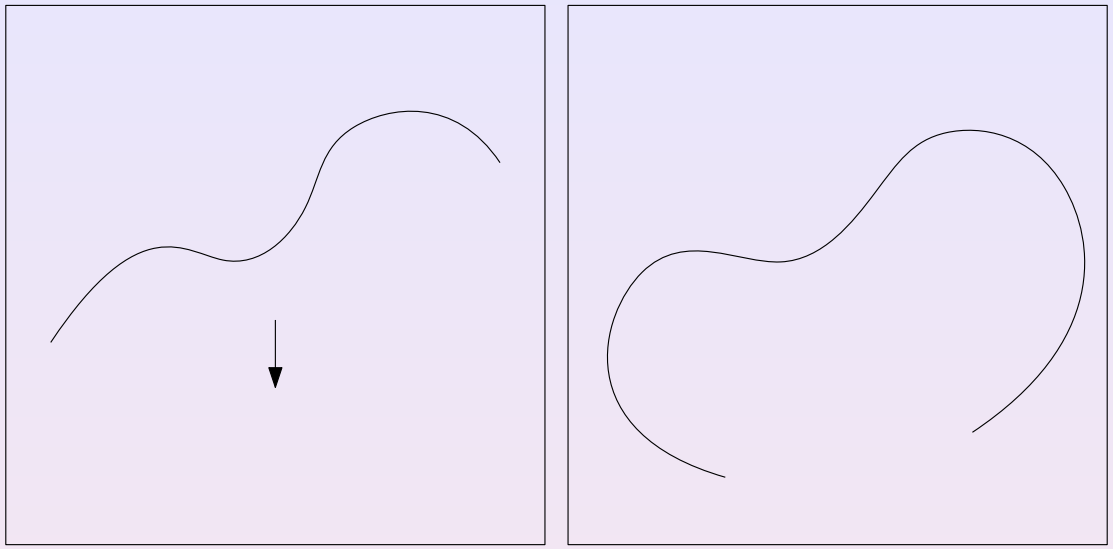
Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, a continuous mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is called an encoding map of $(\mathbb{R}^n, M)$, if restricted on M, $\varphi|_M : M \to \varphi(M) \subset \mathbb{R}^m$ is homeomorphic.

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 $(\mathbb{R}^n, M)$,

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

Encodable Condition



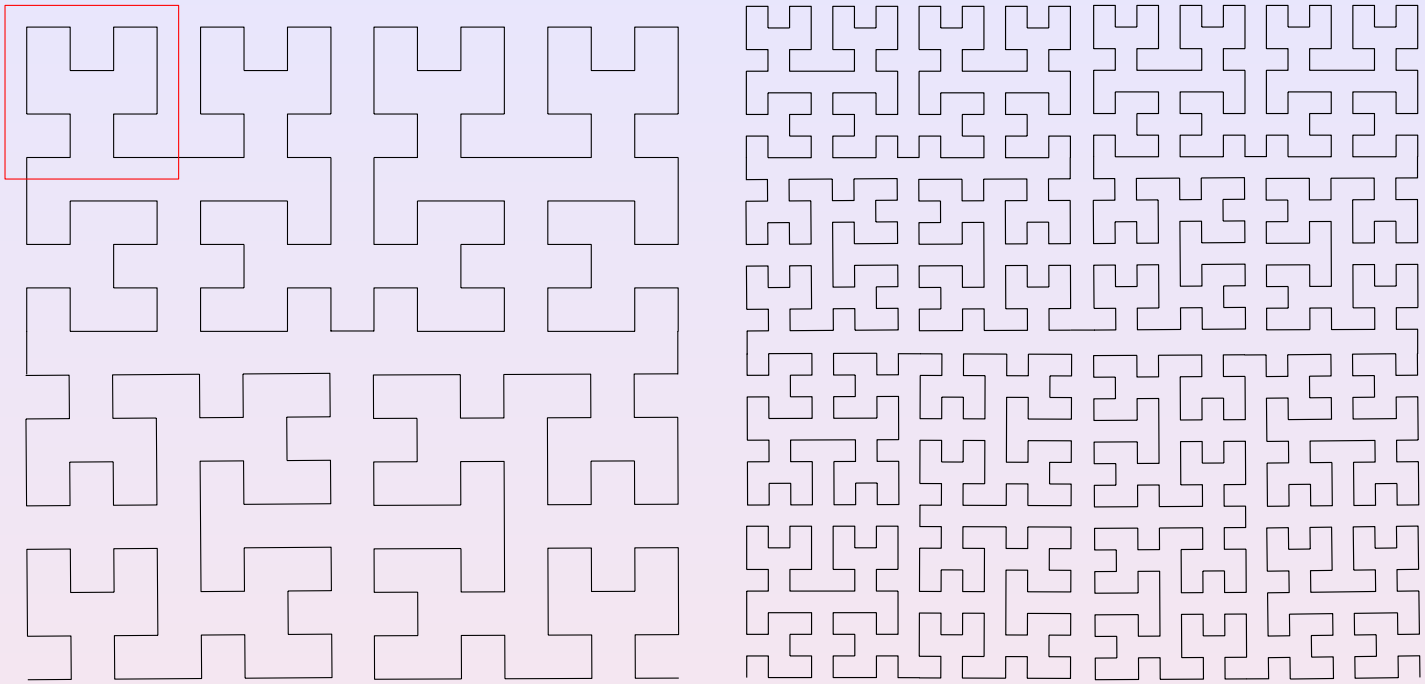
Lemma

Suppose a n dimensional manifold M is embedded in $\mathbb{R}^{n+1}$,

$$M \xrightarrow{G} \mathbb{S}^n \xrightarrow{p} \mathbb{RP}^n$$

where $G : M \to \mathbb{S}^n$ *is the Gauss map,* $\mathbb{RP}^n$ *is the real projective space, if* $p \circ G(M)$ *covers the whole* $\mathbb{RP}^n$*, then M is not linear rectifiable.*

Representation Limitation Theorem



C_1 Peano curve C_2 Peano curve

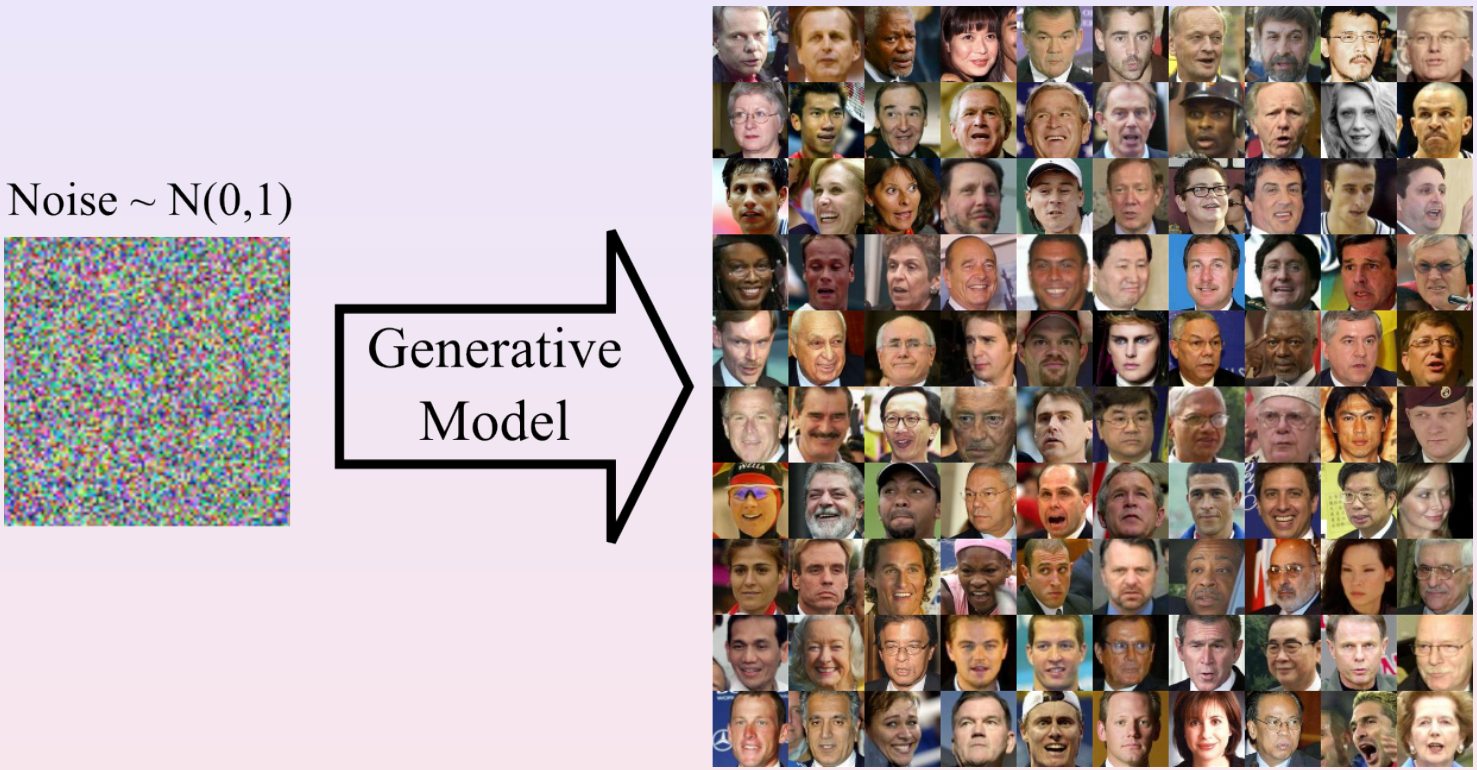
Figure: $\mathcal{N}(\mathbb{R}^2, C_n) \geq 4^{n+1}$

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.

How does DL control the probability distribution?

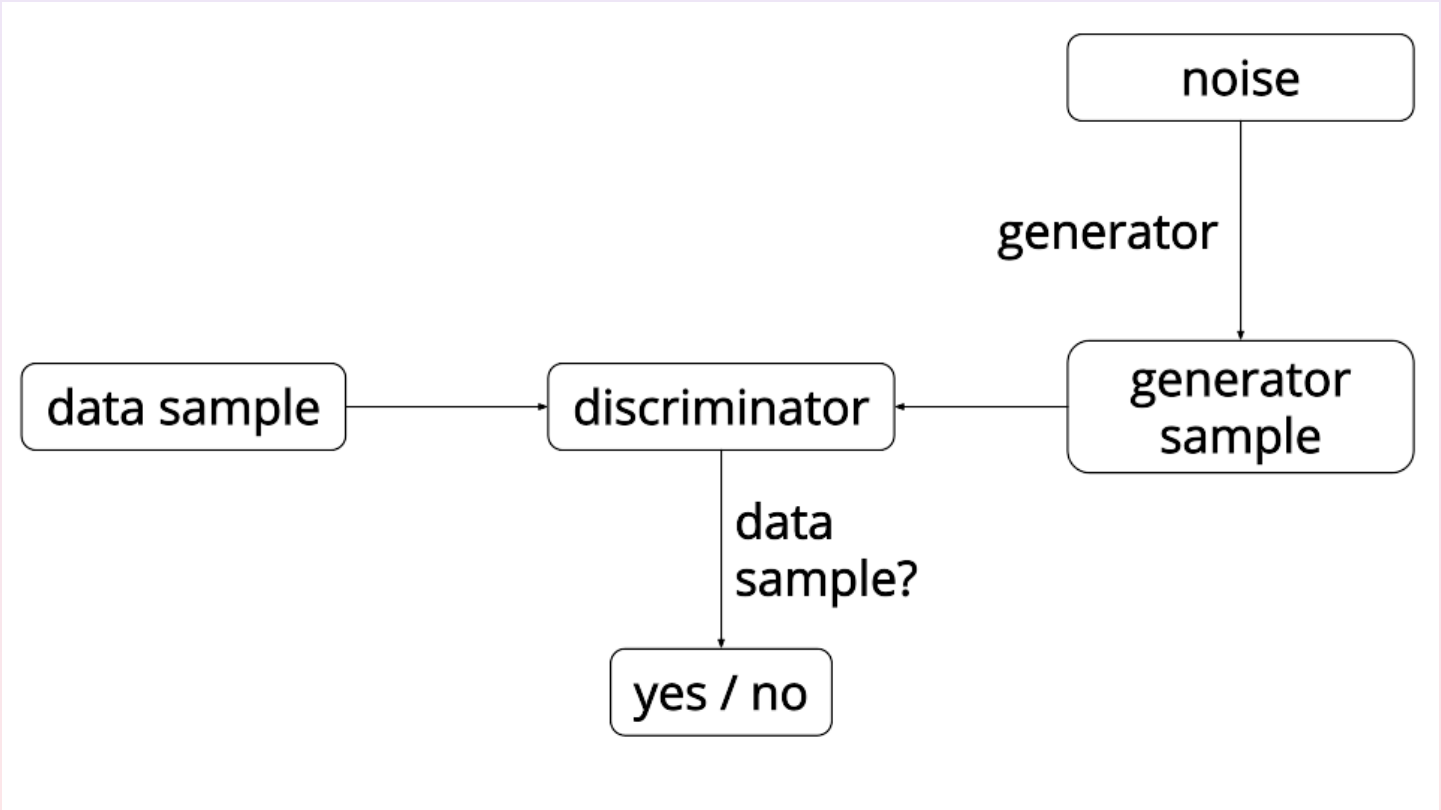
Generative Model



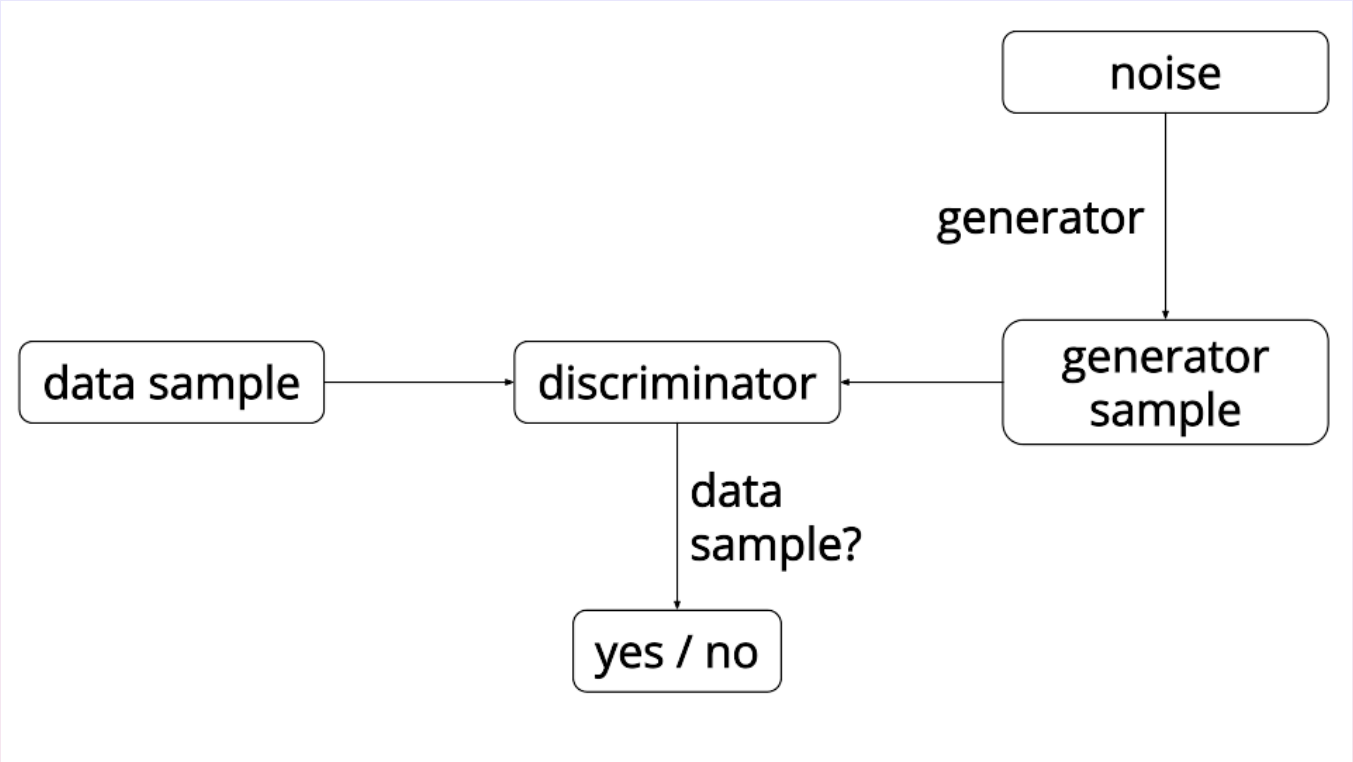
A generative model converts a white noise into a facial image.

GAN Overview

The analogy that is often used here is that the generator is like a forger trying to produce some counterfeit material, and the discriminator is like the police trying to detect the forged items.



GAN Overview



Merits

1. Automatic generate samples, the requirement for the data samples is reduced;
2. Data sample distribution can be arbitrary, without closed form expression.

GAN Overview

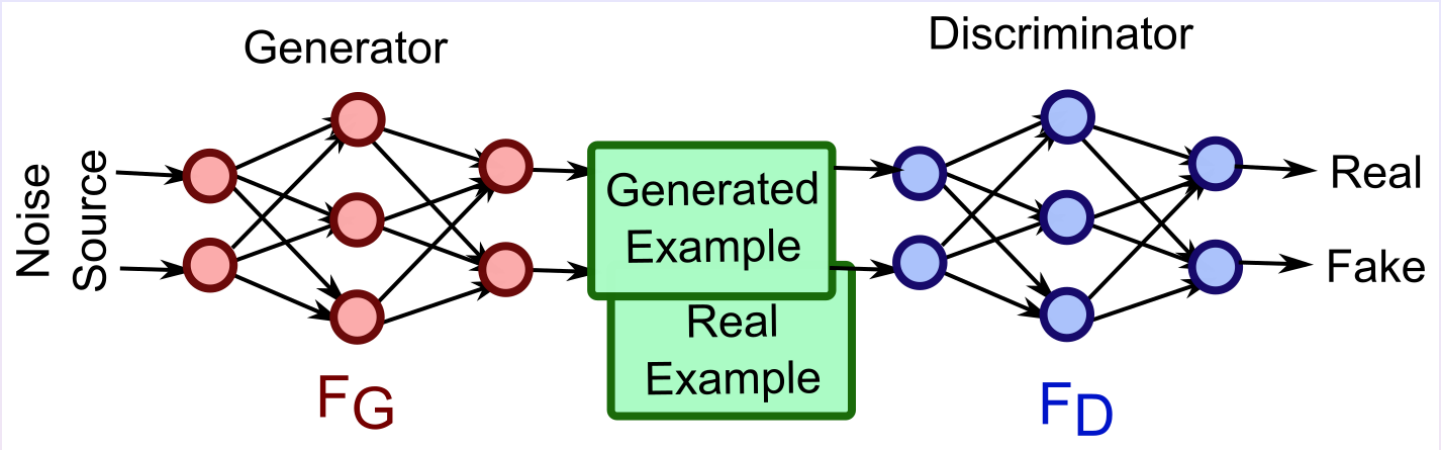


Figure: GAN DNN model.

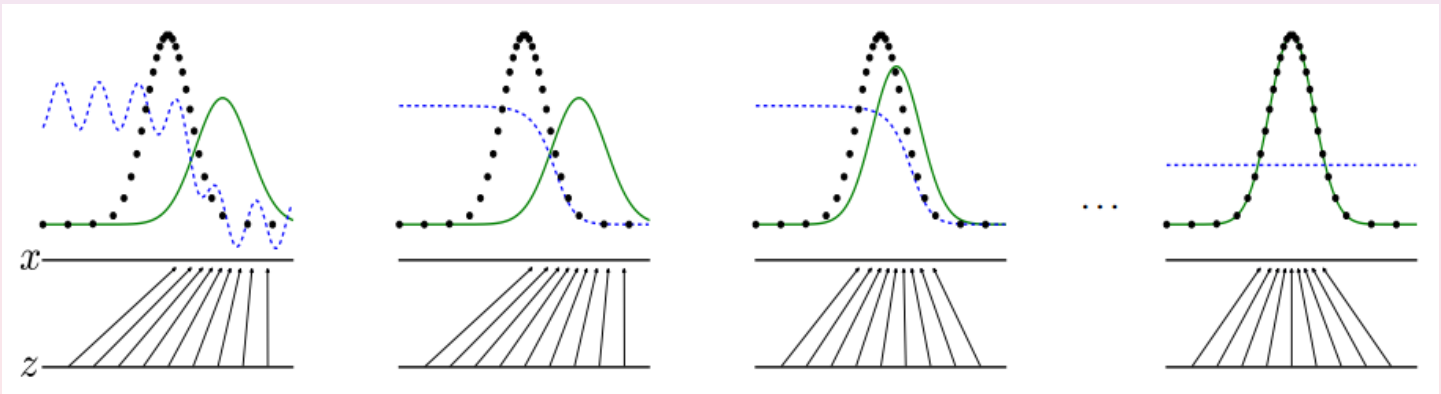
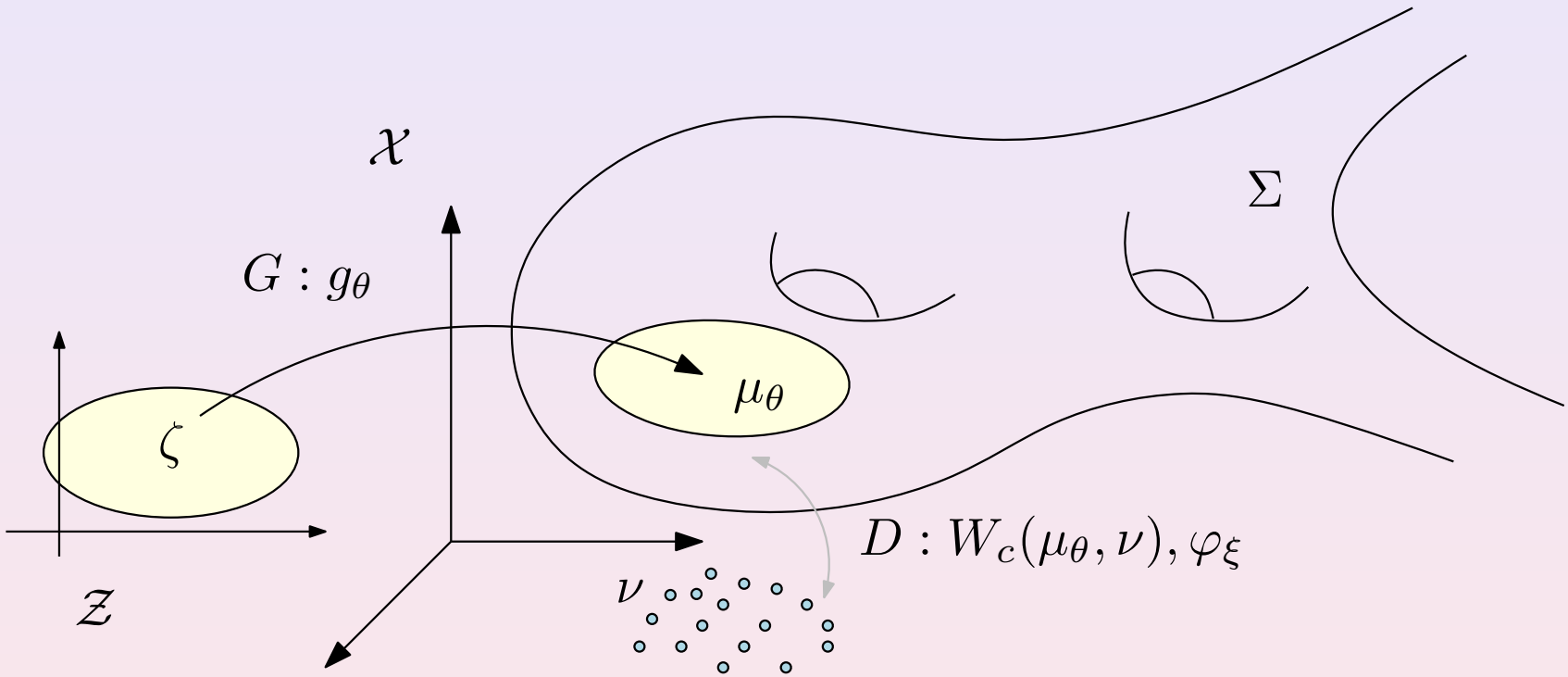


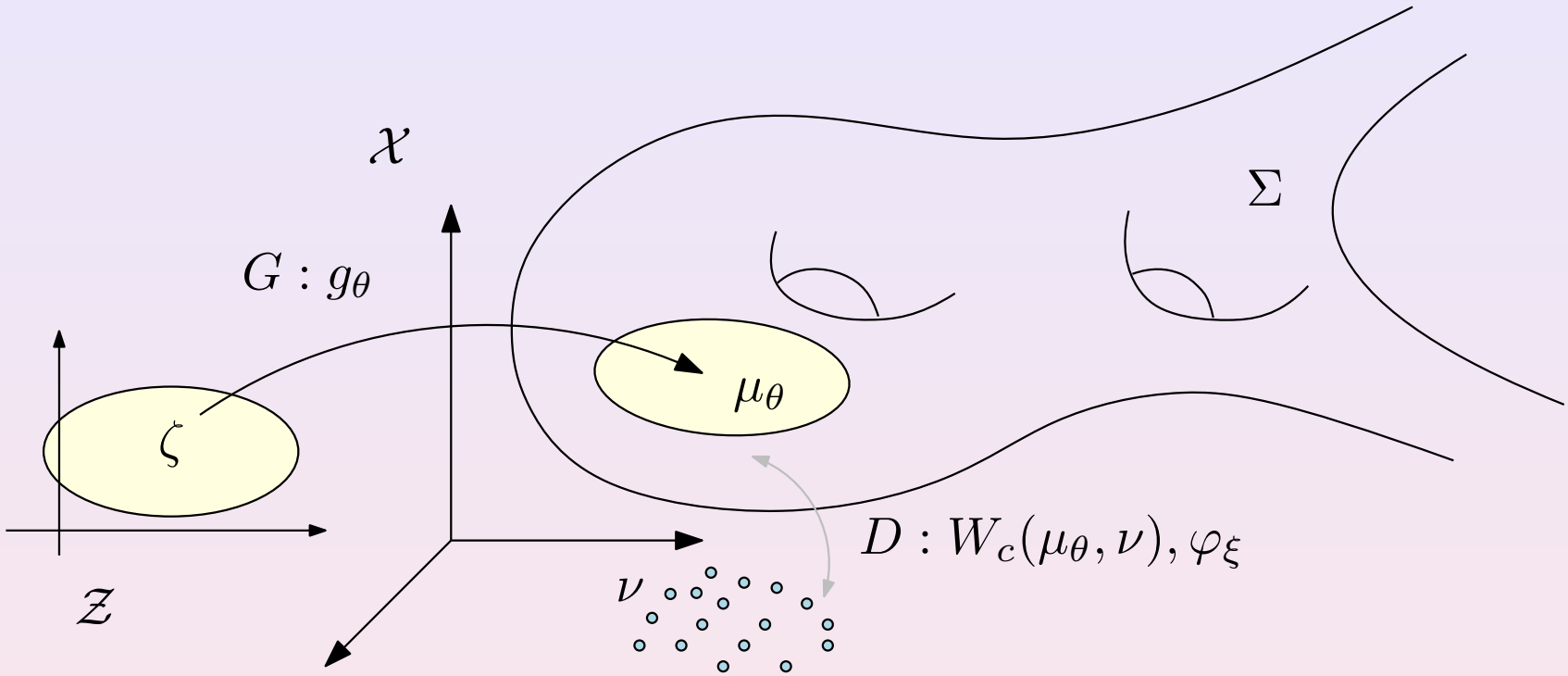
Figure: GAN learning process.

Wasserstein GAN Model



$\mathscr{X}$-image space; Σ-supporting manifold; $\mathscr{Z}$-latent space;

Wasserstein GAN Model

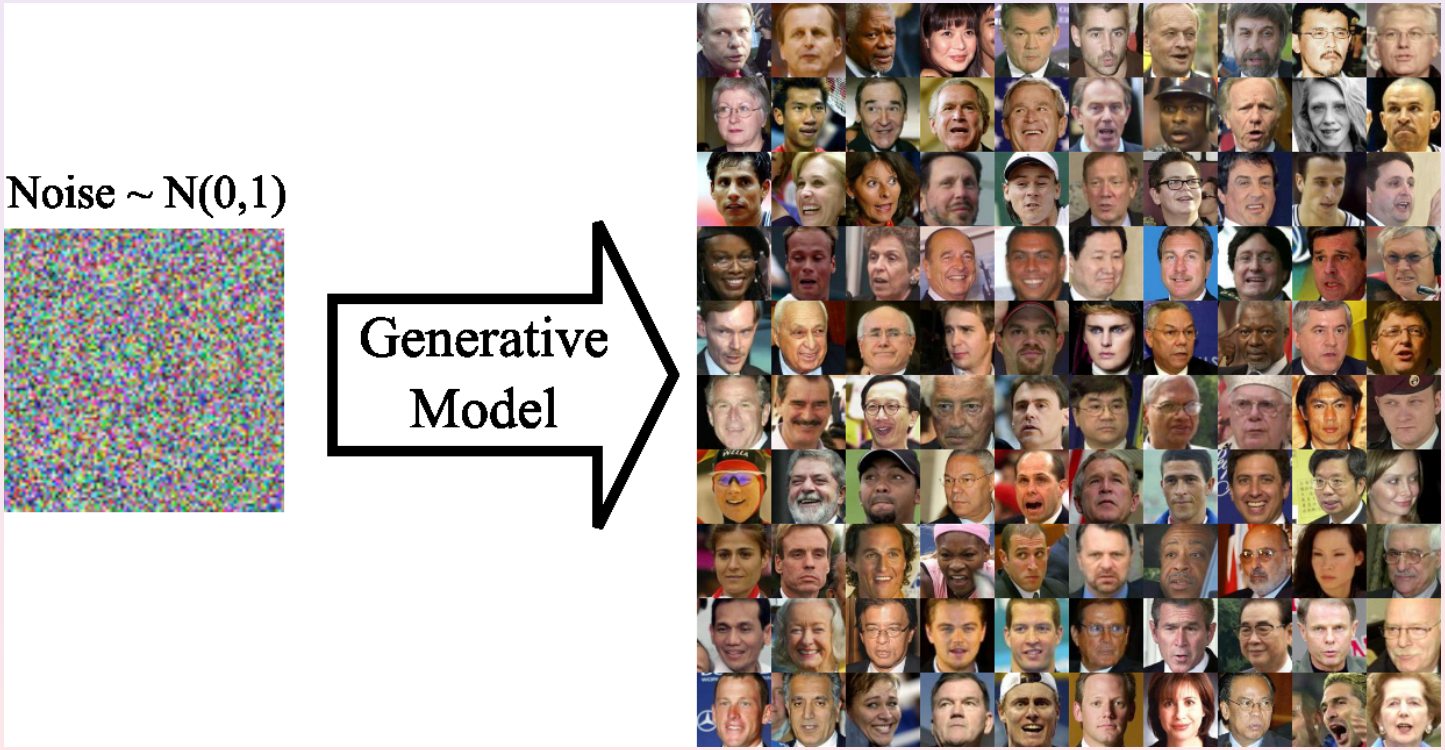


ν-training data distribution; ζ-uniform distribution;
$\mu_\theta = g_{\theta\#}\zeta$-generated distribution; G - generator computes g_θ;
D -discriminator, measures the distance between ν and μ_θ,
$W_c(\mu_\theta, \nu)$.

Generative Model

Generative Model

$G : \mathscr{Z} \to \mathscr{X}$ maps a fixed probability distribution ζ to the training data probability distribution ν.



Overview

Wasserstein Space

Given a Riemannian manifold M, all the probability distributions on M form an infinite dimensional manifold Wasserstein space $\mathscr{W}(M)$, the distance between two probability distributions is given by the so-called Wasserstein distance.

Optimal Mass Transportation

Given two probability measures $\mu, \nu \in \mathscr{W}(M)$, there is a unique optimal mass transportation map $T : M \to M$, φ maps μ to ν with the minimal transportation cost. The transportation cost of the optimal transportation map is the Wasserstein distance between μ and ν.

Definition (Measure-Preserving Mapping)

Given two bounded domains in $\mathbb{R}^n$ with probability measures (X,μ) and (Y,ν), with equal total measure $\mu(X)=\nu(Y)$, a transportation mapping $T: X \to Y$ is measure-preserving, if for any measurable set $B \subset Y$,

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

and denoted as $T_{\#}\mu = \nu$.

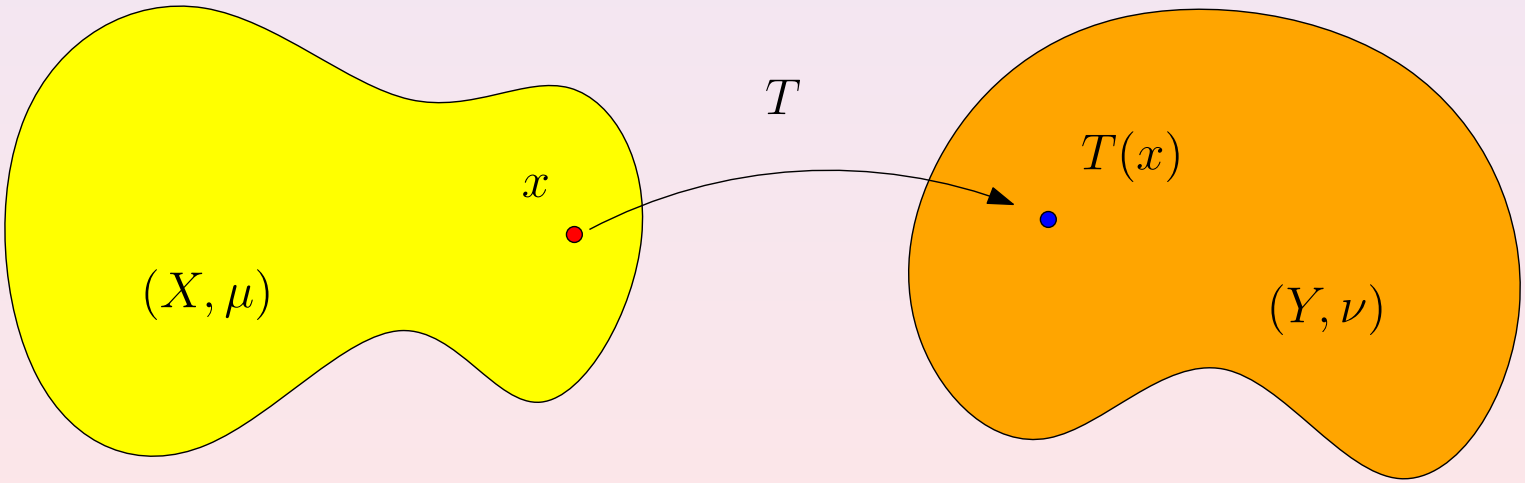
Suppose T is a smooth map, then measure-preserving condition is equivalent to the Jacobian equation $\mu(x)dx = \nu(y)dy$

$$det(DT) = \frac{\mu(x)}{\nu \circ T(x)}.$$

Definition (Transportation Cost)

Suppose the cost of moving a unit mass from point x to point y is $c(x,y)$, for a transportation map $T:(X,\mu)\to(Y,\nu)$, the total transportation cost is

$$\mathscr{C}(T)=\int_X c(\mathbf{x},T(\mathbf{x}))d\mu(\mathbf{x}).$$



Cost Function $c(x, y)$

The cost of moving a unit mass from point x to point y.

$$Monge(1781) : c(x, y) = |x - y|.$$

This is the natural cost function. Other cost functions include

$$\begin{aligned} c(x, y) &= |x - y|^p, p \neq 0 \\ c(x, y) &= -\log|x - y| \\ c(x, y) &= \sqrt{\varepsilon + |x - y|^2}, \varepsilon > 0 \end{aligned}$$

Any function can be cost function. It can be negative.

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

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

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

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

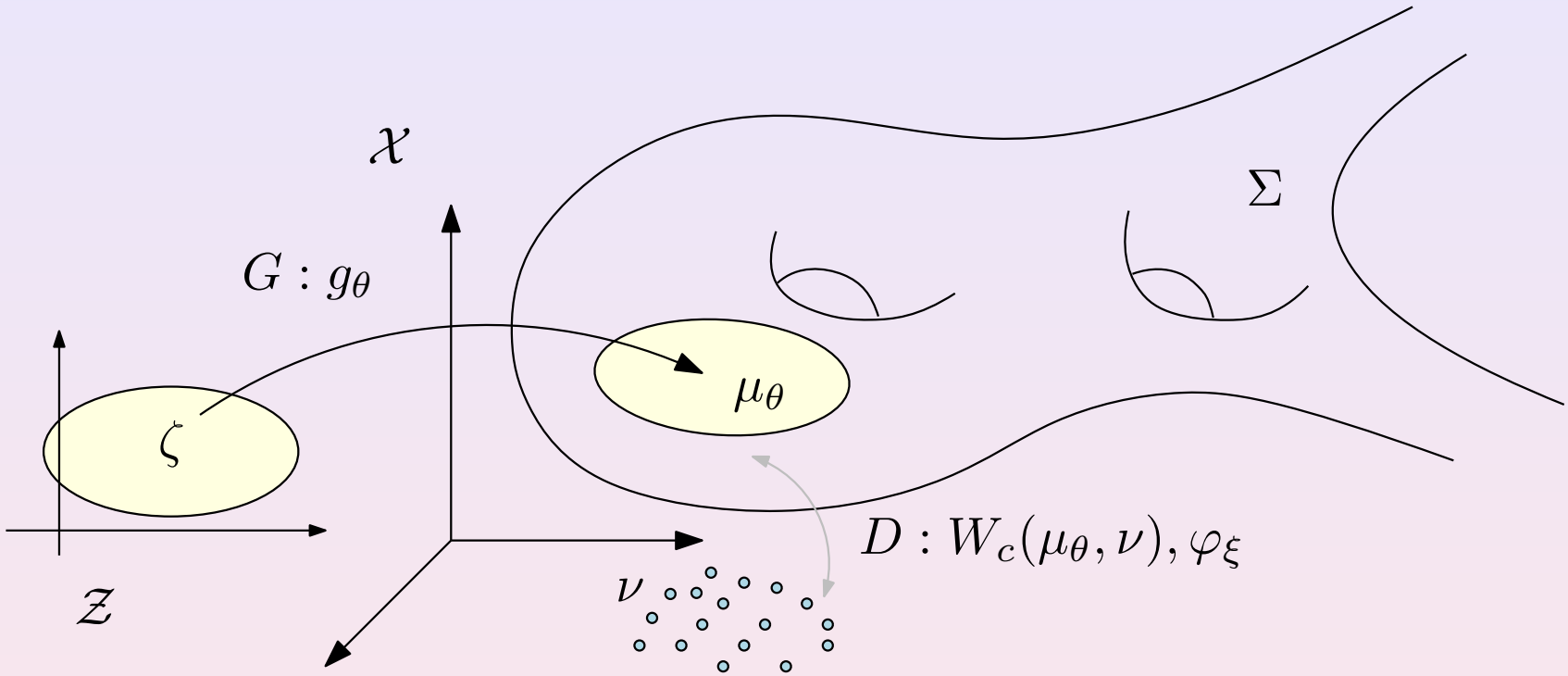
Problem (Kantorovich Dual)

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.

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,
$W_c(\mu_\theta, \nu)$.

OMT view of WGAN

From the optimal transportation point of view, Wasserstein GAN performs the following tasks:

- The discriminator: computes the Wassersteind distance using Kantorovich Dual formula:

$$W_c(\mu_\theta, \nu) = \max_{\varphi_\xi} \int_X \varphi_\xi(x) d\mu_\theta(x) + \int_Y \varphi_\xi^c(y) d\nu(y),$$

  namely computes the Kantorovich potential φ;
- The generator: computes a measure-preserving transportation map $g_\theta : \mathscr{Z} \to \mathscr{X}$, s.t. $g_{\theta\#}\zeta = \mu_\theta = \nu$.
- The WGAN model: min-max optimization

$$\min_\theta \max_\xi \int_X \varphi_\xi \circ g_\theta(z) d\zeta(z) + \int_Y \varphi_\xi^c(y) d\nu(y)$$

OMT view of WGAN

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.

Brenier's Approach

Theorem (Brenier)

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

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

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

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

Kantorovich and Brenier potentials

Theorem

If the distance function $c(x,y) = h(x-y)$, where $h : \mathbb{R} \to \mathbb{R}$ is a strictly convex function, the Kantorovich potential $\varphi : X \to \mathbb{R}$ gives the optimal mass transportation map directly:

$$T(\mathbf{x}) = \mathbf{x} - (\nabla\varphi)^{-1}(\nabla\varphi(\mathbf{x}))$$

Corollary

Suppose $c(x,y) = \frac{1}{2}|x-y|^2$, then Kantorovich potential and Brenier potential satisfy the relation

$$u(\mathbf{x}) = \frac{1}{2}|\mathbf{x}|^2 - \varphi(\mathbf{x}).$$

OMT view of WGAN

L^2 case

The discriminator computes the Kantorovich potential φ; the generator G computes the optimal transportation map, $T = \nabla u$, where u is the Brenier potential; The Brenier potential equals to

$$u = \frac{1}{2}|x|^2 - \varphi(x).$$

Hence, in theory:

- G can be obtained from the optimal D without training;
- D can be obtained from the optimal G without training;
- The two deep neural networks are redundant;
- The competition between D and G is unnecessary.

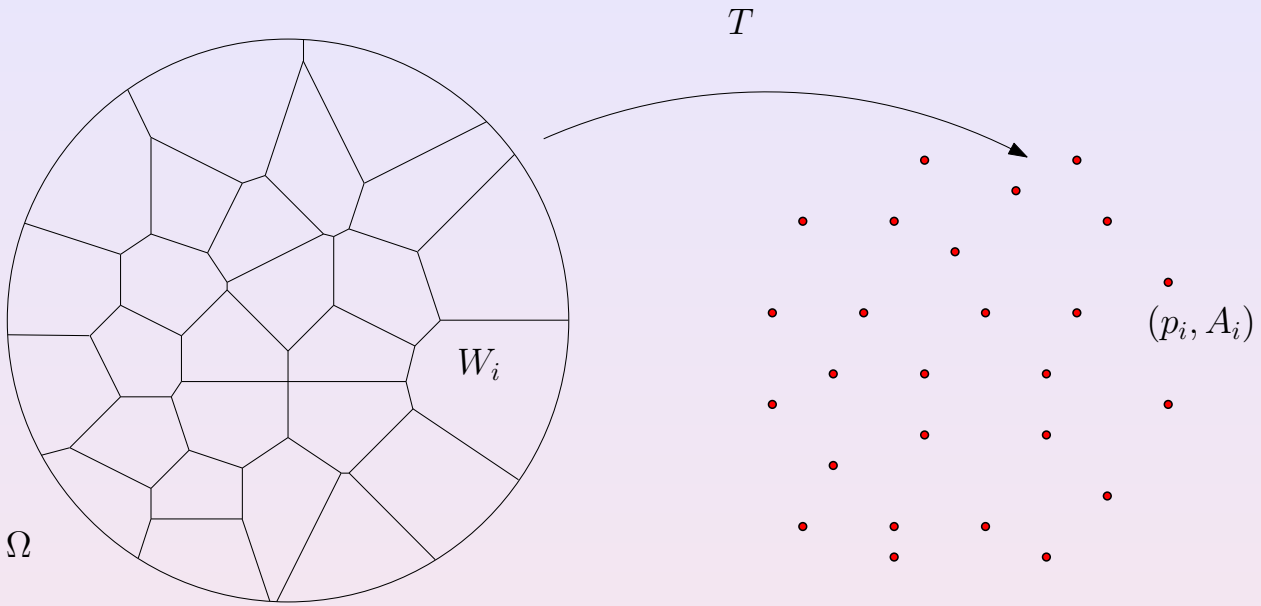
Empirical Distribution

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

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

in general $\nu_i = 1/n$.

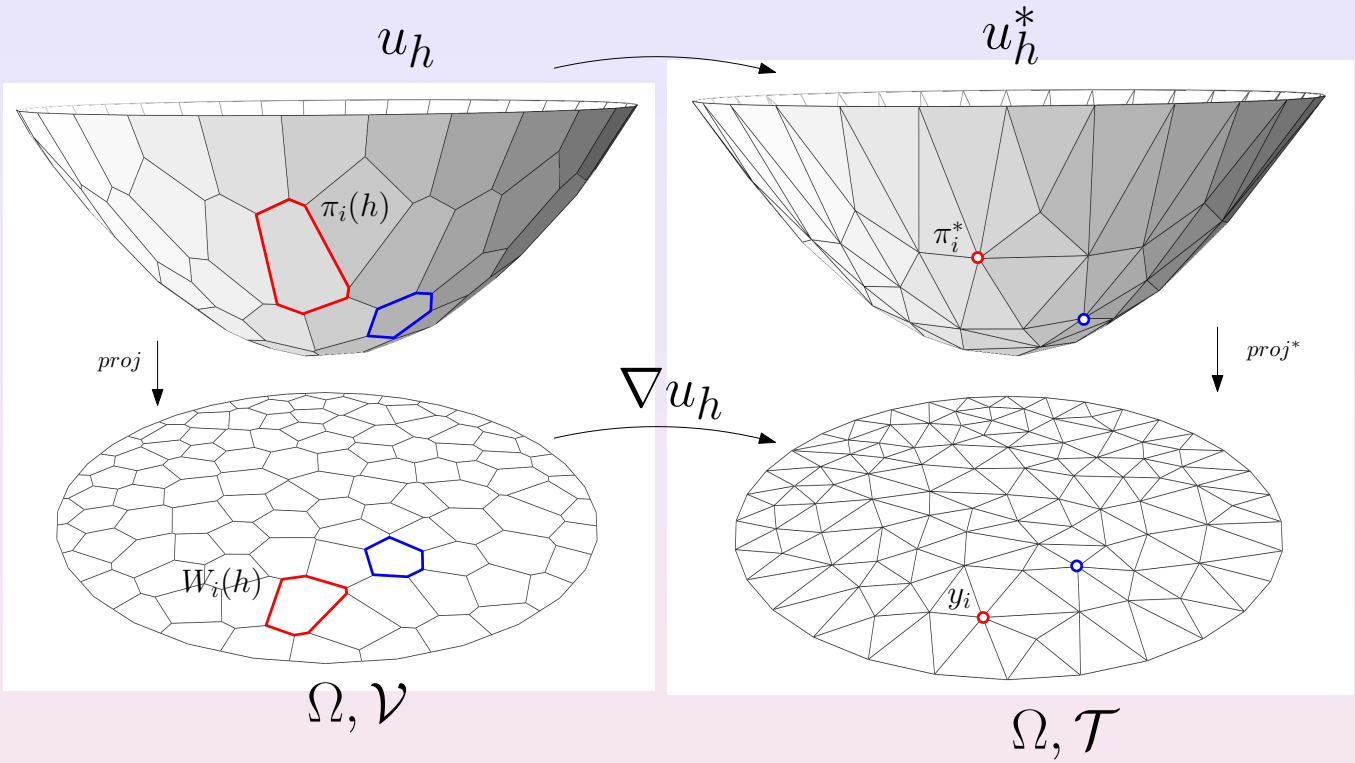
Semi-discrete Optimal Transportation



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, $\mathcal{W}(h) = \cup_i W_i(h)$;
4. adjust the heights h, such that $\mu(W_i(h)) = \nu_i$.

Power Diagram vs Optimal Transport Map

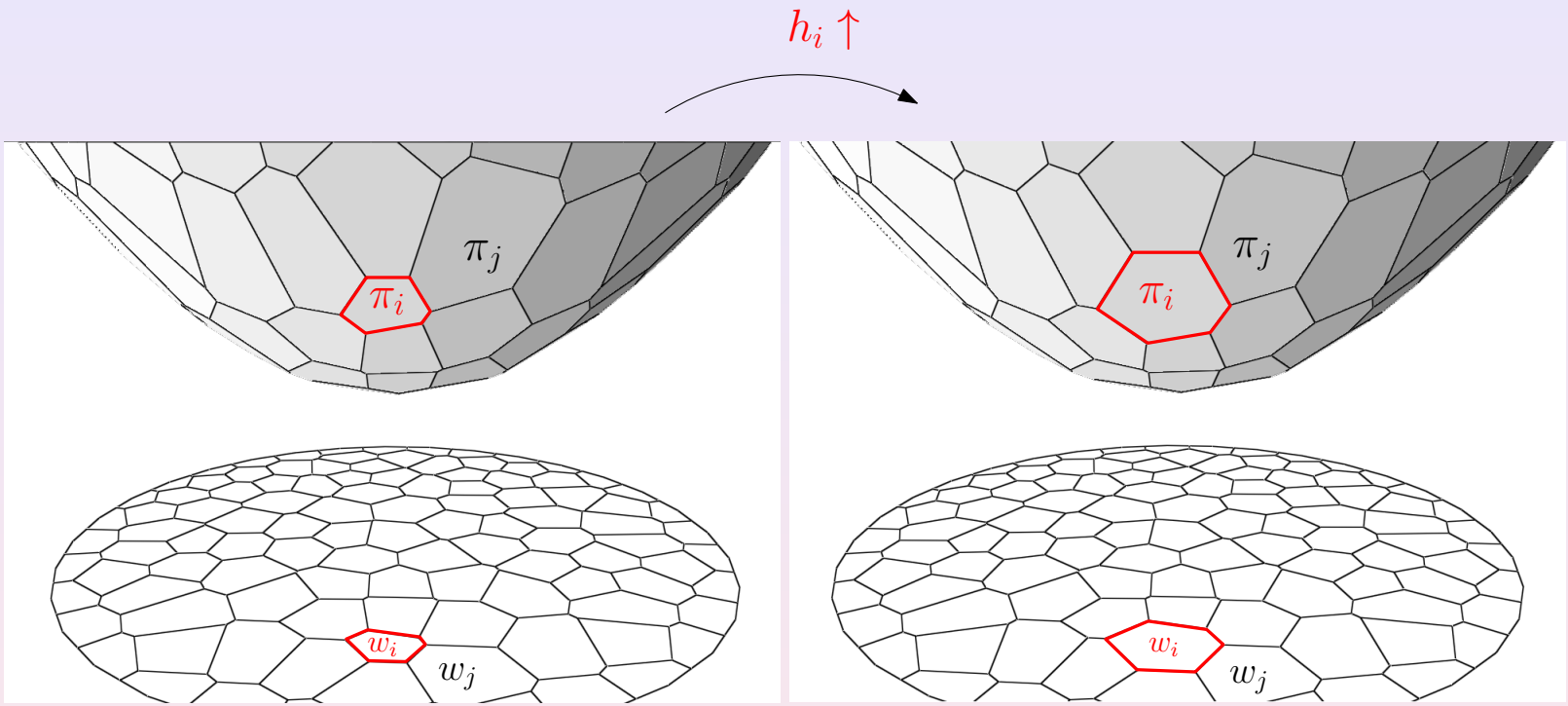


Figure: Variation of the μ-volume of top-dimensional cells.

Adjust the height of each hyper-plane, such that $\mu(W_i(h)) = \nu_i$.

Convex Geometry

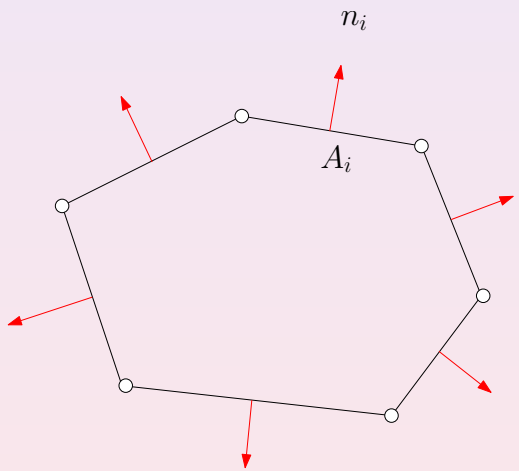
Minkowski problem - 2D Case

Example

A convex polygon P in $\mathbb{R}^2$ is determined by its edge lengths A_i and the unit normal vectors $\mathbf{n}_i$.

Take any $\mathbf{u} \in \mathbb{R}^2$ and project P to $\mathbf{u}$, then $\langle \sum_i A_i \mathbf{n}_i, \mathbf{u} \rangle = \mathbf{0}$, therefore

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



Minkowski problem - General Case

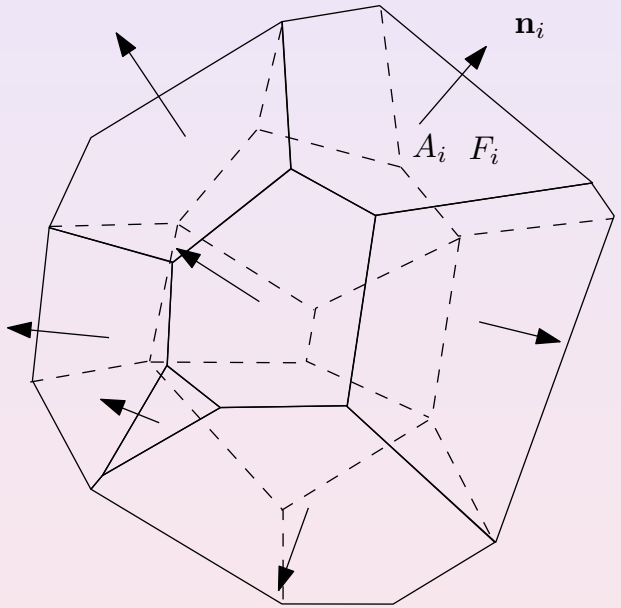
Minkowski Problem

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

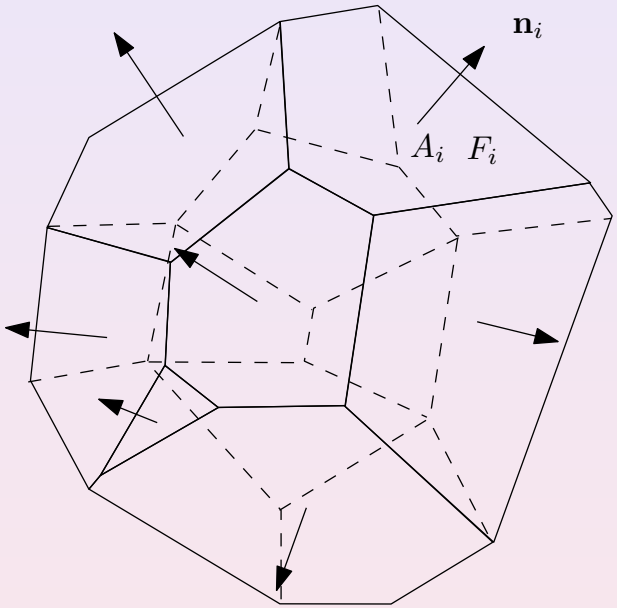
1. $area(F_i) = A_i$,
2. $\mathbf{n}_i \perp F_i$.



Minkowski problem - General Case

Theorem (Minkowski)

P exists and is unique up to translations.



Brunn-Minkowski inequality

Theorem (Brunn-Minkowski)

For every pair of nonempty compact subsets A and B of $\mathbb{R}^n$ and every $0 \leq t \leq 1$,

$$[Vol(tA \oplus (1-t)B)]^{\frac{1}{n}} \geq t[vol(A)]^{\frac{1}{n}} + (1-t)[vol(B)]^{\frac{1}{n}}.$$

For convex sets A and B, the inequality is strick for $0 < t < 1$ unless A and B are homothetic i.e. are equal up to translation and dilation.

Alexandrov Theorem

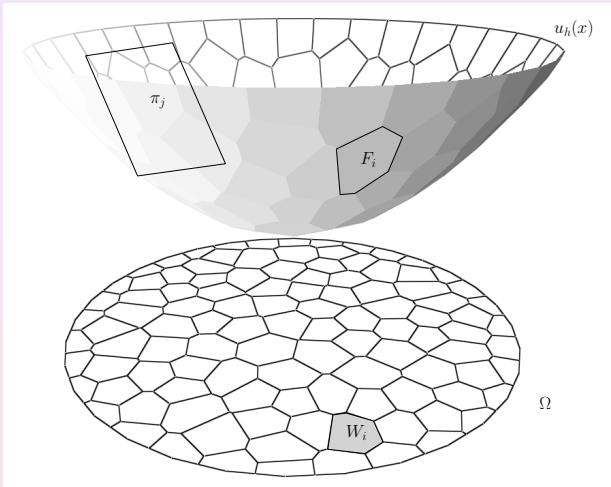
Theorem (Alexandrov 1950)

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

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

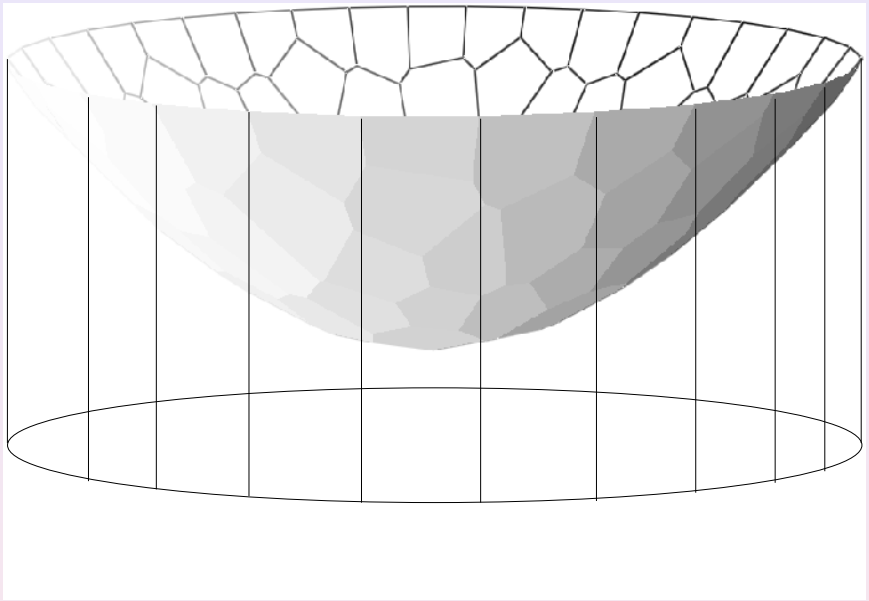
where $w_i(\eta) = \mu(W_i(\eta) \cap \Omega)$ is the μ-volume of the cell.

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

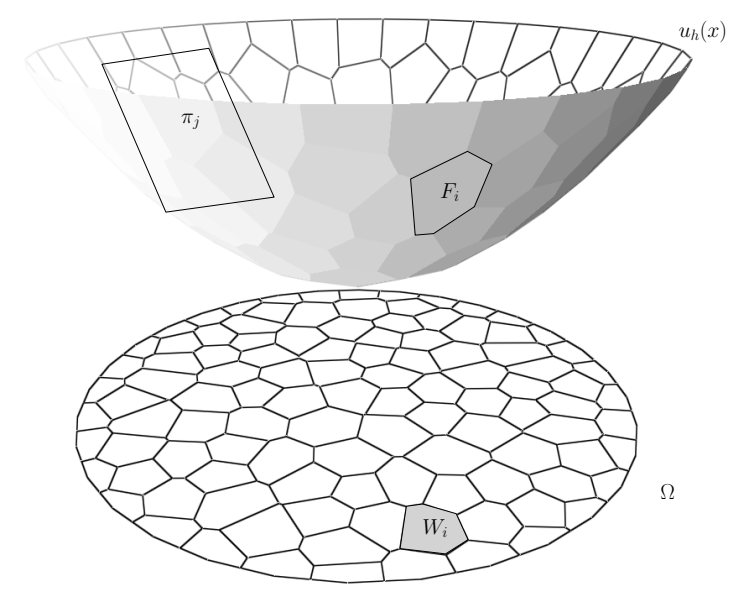
Accepted by Asian Journal of Mathematics (AJM)

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



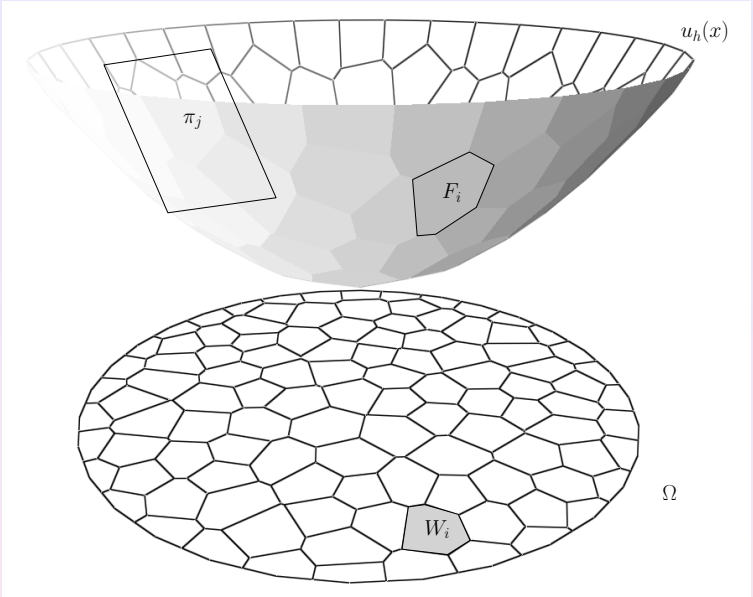
Definition (Alexandrov Potential)

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.

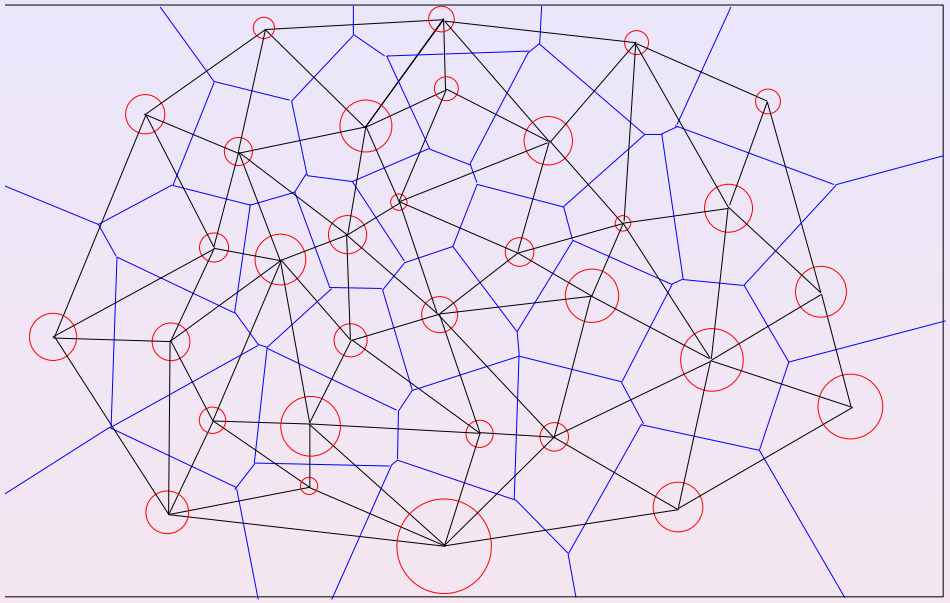
Computational Algorithm



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

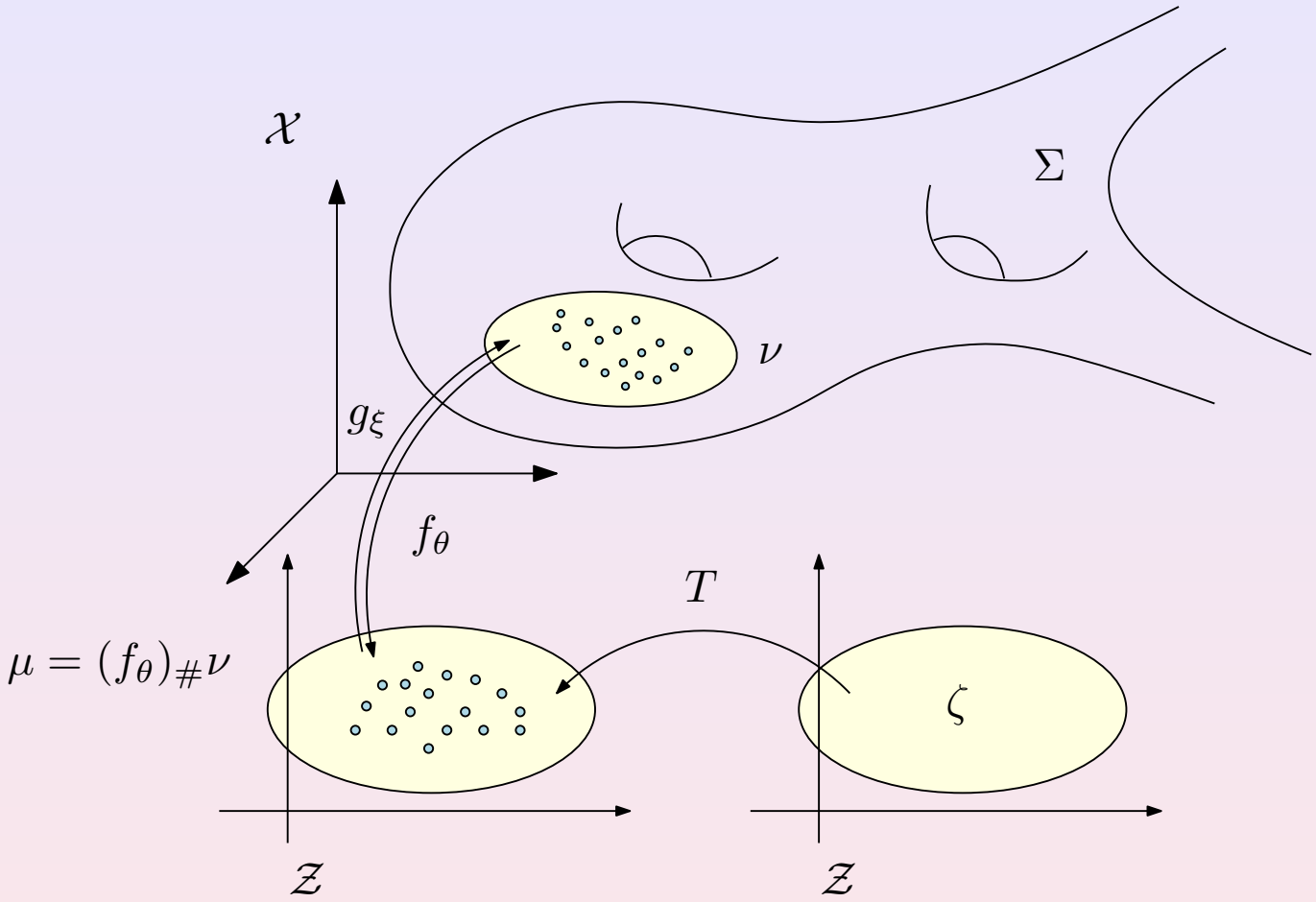


The Hessian of the energy is the length ratios of edge and dual edges,

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

Generative Model

Auntoencoder-OMT



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

Experiments

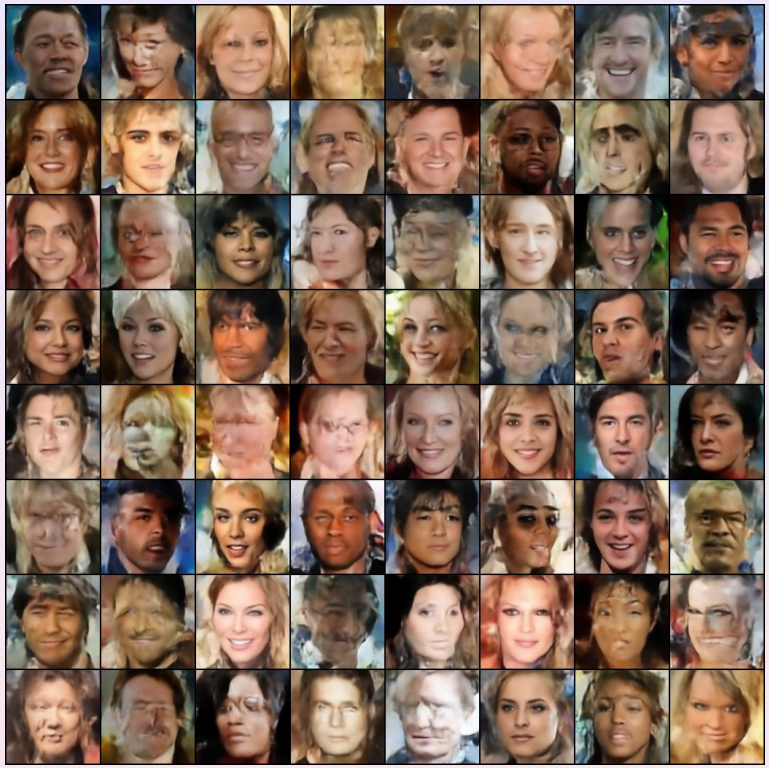
(a) real digits

(b) VAE

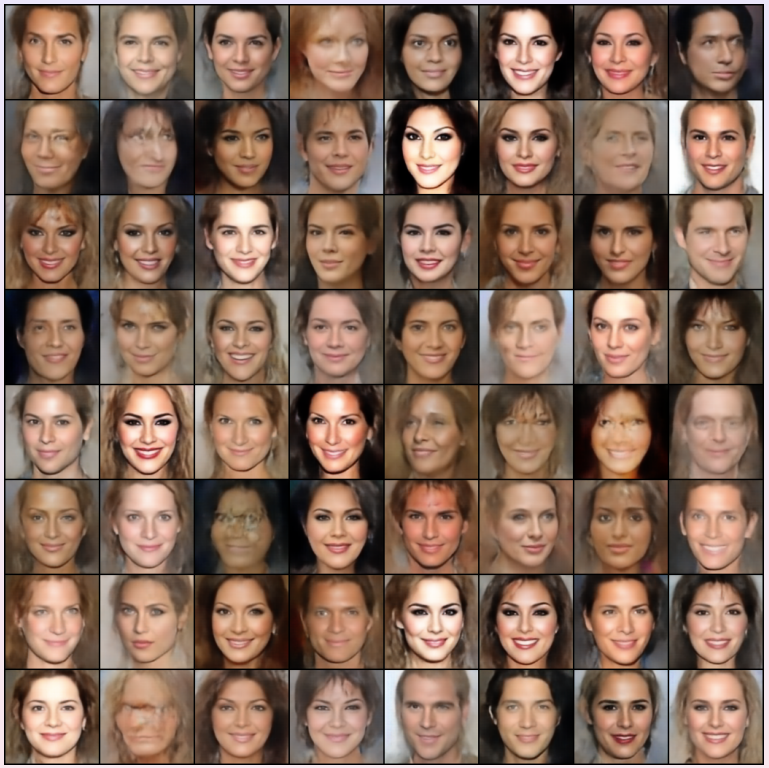
(c) WGAN

(d) AE-OMT

Experiments



(a) VAE



(d) AE-OMT

Conclusion

This work introduces a geometric understanding of deep learning:

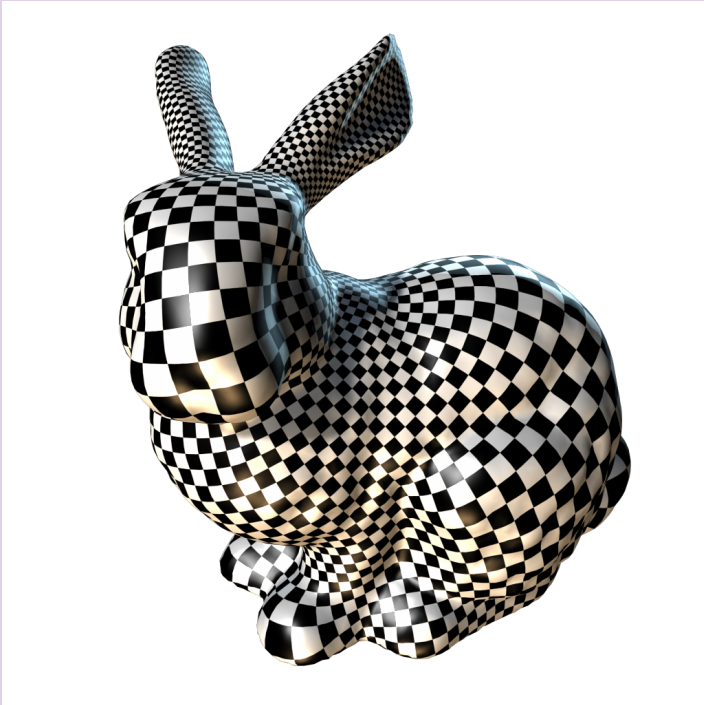
- The manifold distribution assumption and the clustering assumption.
- Network complexity; manifold complexity.
- Geometric theory of optimal mass transportation.

References

- Na Lei, Zhongxuan Luo, Shing-Tung Yau and Xianfeng Gu, “Geometric Understanding of Deep Learning”, arXiv:1805.10451
- Na Lei, Kehua Su, Li Cui, Shing-Tung Yau and Xianfeng Gu, “A Geometric View of Optimal Transportation and Generative Model”, arXiv:1710.54888
- Xianfeng Gu, Feng Luo, Jian Sun and Shing-Tung Yau, Variational Principles for Minkowski Type Problems, Discrete Optimal Transport, and Discrete Monge-Ampere Equations, Vol. 20, No. 2, pp. 383-398, Asian Journal of Mathematics (AJM), April 2016.

Thanks

For more information, please email to gu@cs.stonybrook.edu.



Thank you!