Mathematical Foundations of Computation, Deep Learning, and Quantum TDA

Yuan YAO

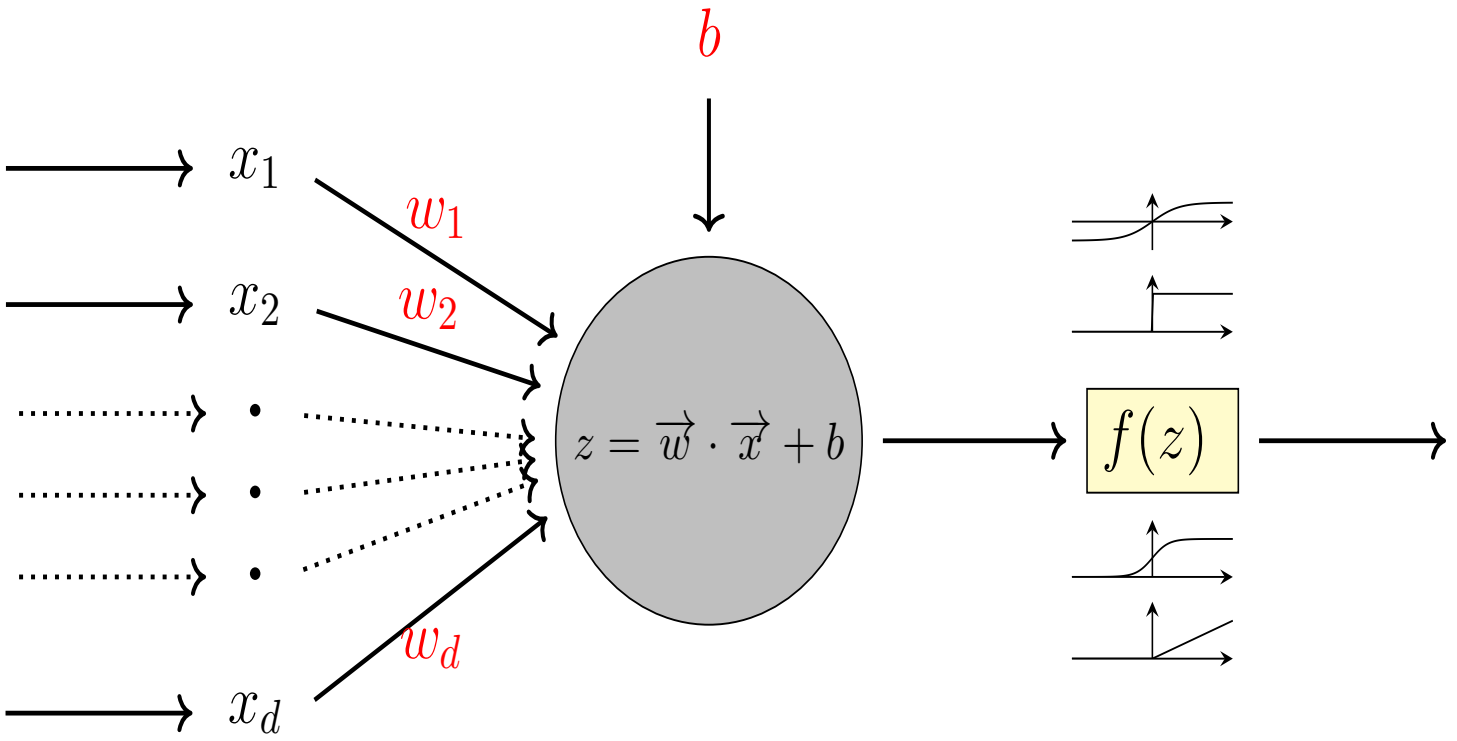
HKUST

Outline

- Perceptrons and Finite Halting
- **Minsky-Papert (1969)** Model:
  - locality/sparsity is a fundamental limit of neural networks
  - XOR and Topological computation
- Deep Neural Networks
  - complexity and the curse-of-dimensionality (**Poggio** et al. 2017)
  - Geometric group invariants in deep networks (**Mallat** et al. 2012; **Bolcskei** et al. 2017)
- **Blum-Shub-Smale (1989)** Model of Real Computation
  - Condition number is a fundamental limit of topology learning with finite samples
- Topological Data Analysis
  - Computing robust topological invariants to scales and the curse-of-dimensionality
- Quantum algorithms for topological data analysis
  - Polynomial complexity (**Lloyd** et al. 2016)
  - Demonstration of 3-point data by 6-photon quantum computer (**Huang** et al. 2018)

Perceptron: single-layer

- Invented by Frank Rosenblatt (1957)



The Perceptron Algorithm

$$\ell(w) = -\sum_{i \in \mathcal{M}_w} y_i \langle w, \mathbf{x}_i \rangle, \quad \mathcal{M}_w = \{i : y_i \langle \mathbf{x}_i, w \rangle < 0, y_i \in \{-1, 1\}\}.$$

The Perceptron Algorithm is a *Stochastic Gradient Descent* method:

$$\begin{aligned} w_{t+1} &= w_t - \eta_t \nabla_i \ell(w) \\ &= \begin{cases} w_t - \eta_t y_i \mathbf{x}_i, & \text{if } y_i w_t^T \mathbf{x}_i < 0, \\ w_t, & \text{otherwise.} \end{cases} \end{aligned}$$

Finiteness of Stopping Time

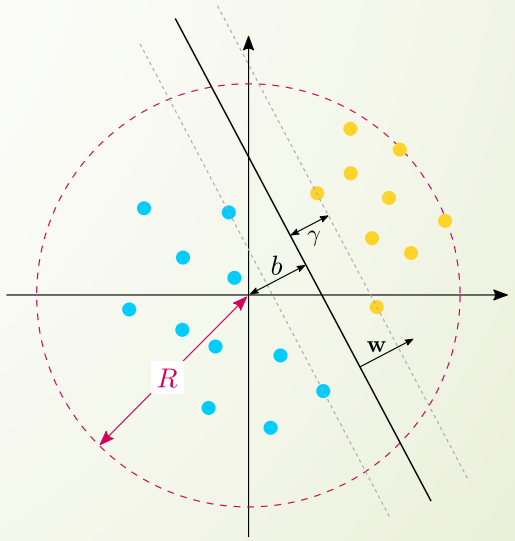
The perceptron convergence theorem was proved by Block (1962) and Novikoff (1962). The following version is based on that in Cristianini and Shawe-Taylor (2000).

Theorem 1 (Block, Novikoff). *Let the training set* $S = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$ *be contained in a sphere of radius* R *about the origin. Assume the dataset to be linearly separable, and let* $\mathbf{w}_{\text{opt}}$, $\|\mathbf{w}_{\text{opt}}\| = 1$, *define the hyperplane separating the samples, having functional margin* $\gamma > 0$. *We initialise the normal vector as* $\mathbf{w}_0 = \mathbf{0}$. *The number of updates,* k, *of the perceptron algorithms is then bounded by*

$$k \leq \left(\frac{2R}{\gamma}\right)^2. \tag{10}$$

Input ball: $R = \max_i \|\mathbf{x}_i\|.$

Margin: $\gamma := \min_i y_i f(x_i)$

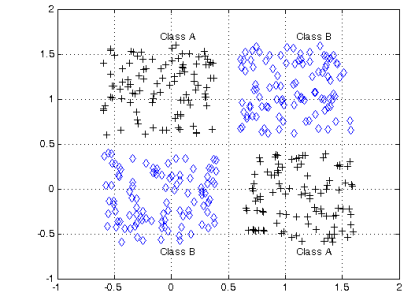


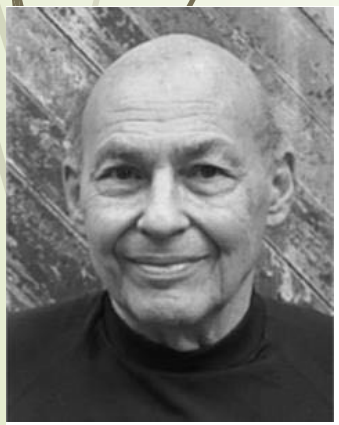
Locality or Sparsity of Computation

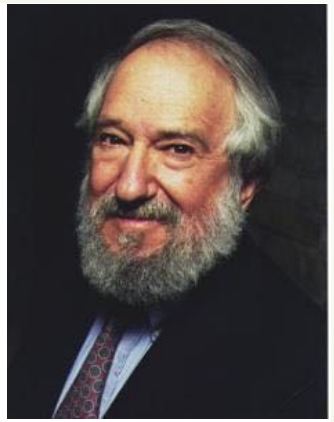
Minsky and Papert, 1969

Perceptron can't do **XOR** classification

Perceptron needs infinite global information to compute **connectivity**





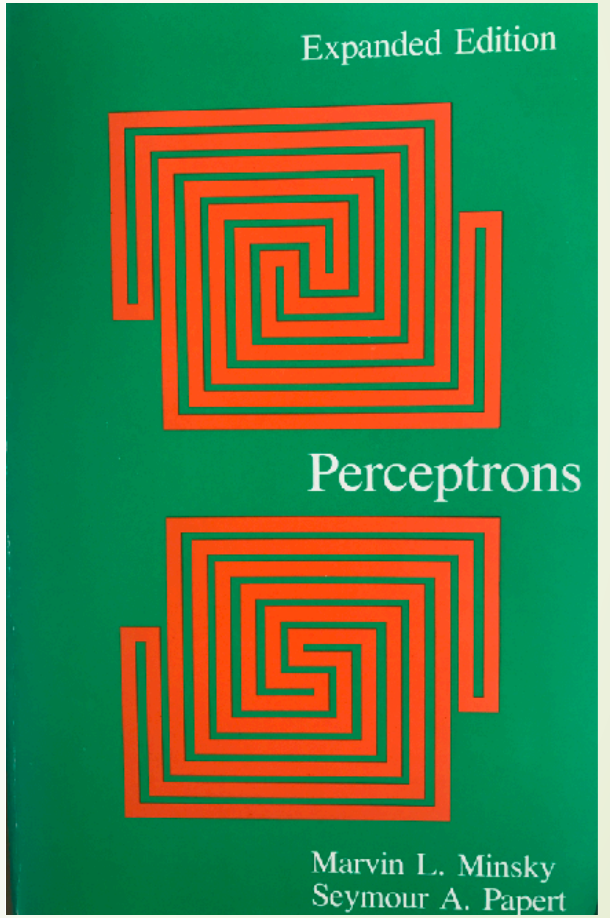


Marvin Minsky Seymour Papert

Locality or **Sparsity** is important:

Locality in time?

Locality in space?



Locality or Sparsity is a fundamental limitation

Definition (Minsky-Papert'1969)

The decision function that $f(X) \in \{1, -1\}$ for $X \subseteq \mathbb{R}^p$ has **order** k, if it can be represented by a superposition of functions whose supports are at most k, i.e. there exists a (possibly of infinite members) family of $\{\phi_\alpha(X) : \mathrm{supp}(\phi_\alpha) \leq k\}$ such that

$$f(X) = \sum_\alpha \phi_\alpha(X)$$

- Minsky-Papert model admits **infinitely many neurons** (wide network) in parallel processing, yet only **sparse** or **local inputs.** Note that it is *not* a Turing model.

Examples of Finite Orders

- $f(X) = [X \text{ is nonempty}]$ has order 1, as $\phi_a(X) = [a \in X]$ and $f(X) = \sum_a \phi_a(X)$.
- $f(X) = [X \text{ is convex}]$ has order 3, as

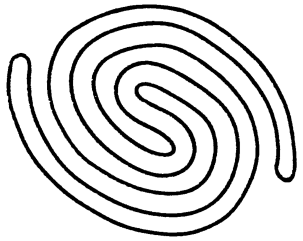
$$f(X) = -\sum_{a,b \in X} [\text{midpoint } ([a,b]) \text{ not in } X]$$

- The only topologically invariant predicates of finite order are functions of the Euler number $E(X)$, which for simplicial complex $X \subseteq \mathbb{R}^2$ is defined as

$$\begin{aligned} E(X) &:= \#(\text{faces } (X)) - \#(\text{edges } (X)) + \#(\text{vertices } (X)) \\ &= \beta_0 - \beta_1 \end{aligned}$$

Connectivity is of infinite order

- Which one of these two figures is connected?



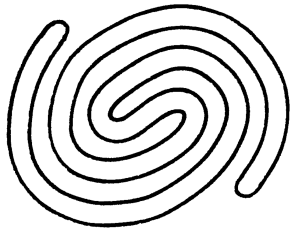


Figure 5.1

Theorem (Minsky-Papert'1969)

The decision function that $f(X) = [X \text{ is connected}]$ *for* $X \subseteq \mathbb{R}^p$ *is* **not of any finite order,** *i.e. for any* $k < \infty$*, there does not exist a (possibly of infinite members) family of* $\{\phi_\alpha(X) : \text{supp}(\phi_\alpha) \leq k\}$ *whose supports are at most* k*, such that*

$$f(X) = \left[\sum_\alpha \phi_\alpha(X) \geq 0\right] \tag{21}$$

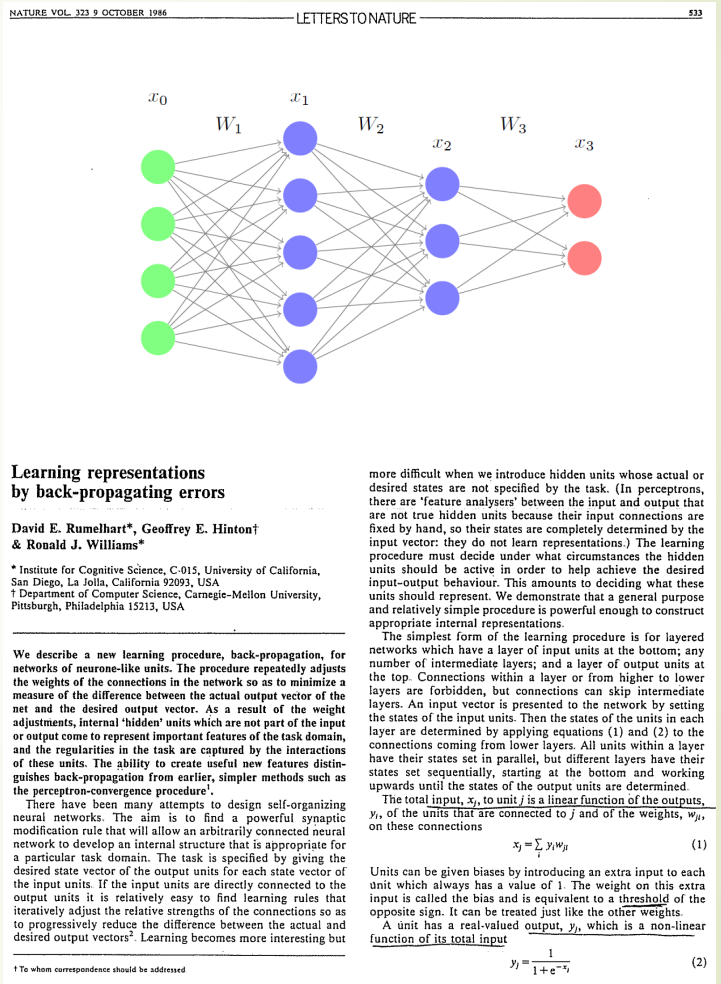
Multilayer Perceptrons (MLP) and Back-Propagation (BP) Algorithms

D.E. Rumelhart, G. Hinton, R.J. Williams (1986)
Learning representations by back-propagating errors, Nature, 323(9): 533-536

BP algorithms as **stochastic gradient descent** algorithms (**Robbins–Monro 1950; Kiefer-Wolfowitz 1951**) with Chain rules of Gradient maps

Deep network may classify **XOR**. Yet **topology**?

We address complexity and geometric invariant properties first.



NATURE VOL. 323 9 OCTOBER 1986 — LETTERS TO NATURE — 533

Learning representations by back-propagating errors

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure[1].

There have been many attempts to design self-organizing neural networks. The aim is to find a powerful synaptic modification rule that will allow an arbitrarily connected neural network to develop an internal structure that is appropriate for a particular task domain. The task is specified by giving the desired state vector of the output units for each state vector of the input units. If the input units are directly connected to the output units it is relatively easy to find learning rules that iteratively adjust the relative strengths of the connections so as to progressively reduce the difference between the actual and desired output vectors[2]. Learning becomes more interesting but more difficult when we introduce hidden units whose actual or desired states are not specified by the task. (In perceptrons, there are 'feature analysers' between the input and output that are not true hidden units because their input connections are fixed by hand, so their states are completely determined by the input vector: they do not learn representations.) The learning procedure must decide under what circumstances the hidden units should be active in order to help achieve the desired input-output behaviour. This amounts to deciding what these units should represent. We demonstrate that a general purpose and relatively simple procedure is powerful enough to construct appropriate internal representations.

The simplest form of the learning procedure is for layered networks which have a layer of input units at the bottom; any number of intermediate layers; and a layer of output units at the top. Connections within a layer or from higher to lower layers are forbidden, but connections can skip intermediate layers. An input vector is presented to the network by setting the states of the input units. Then the states of the units in each layer are determined by applying equations (1) and (2) to the connections coming from lower layers. All units within a layer have their states set in parallel, but different layers have their states set sequentially, starting at the bottom and working upwards until the states of the output units are determined.

The total input, x_j, to unit j is a linear function of the outputs, y_i, of the units that are connected to j and of the weights, w_{ji}, on these connections

$$x_j = \sum_i y_i w_{ji} \quad (1)$$

Units can be given biases by introducing an extra input to each unit which always has a value of 1. The weight on this extra input is called the bias and is equivalent to a threshold of the opposite sign. It can be treated just like the other weights.

A unit has a real-valued output, y_j, which is a non-linear function of its total input

$$y_j = \frac{1}{1+e^{-x_j}} \quad (2)$$

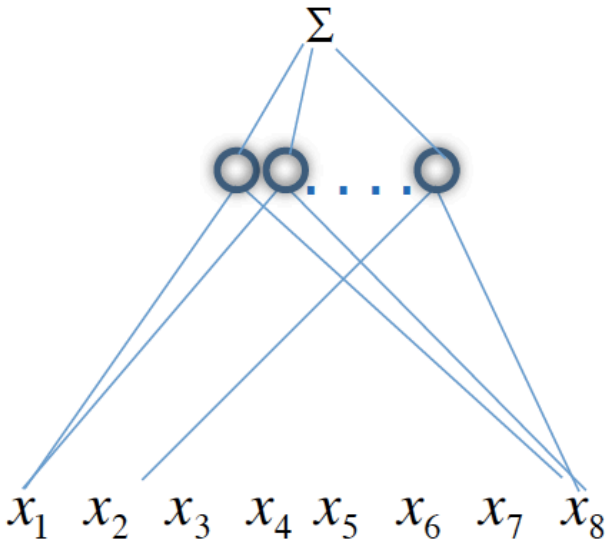
† To whom correspondence should be addressed

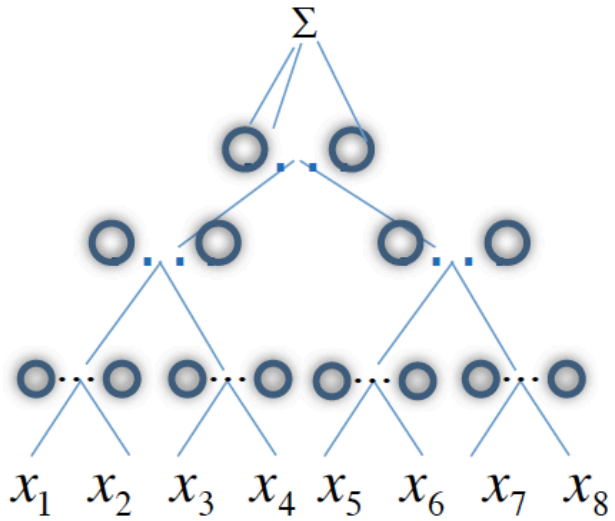
Complexity of Deep Networks and the Curse of Dimensionality

Deep networks can avoid the curse of dimensionality with locality or sparsity.

Deep and shallow networks: universality

Theorem *Shallow, one-hidden layer networks with a nonlinear $\phi(x)$ which is not a polynomial are universal. Arbitrarily deep networks with a nonlinear $\phi(x)$ (including polynomials) are universal.*





$$\phi(x) = \sum_{i=1}^{r} c_i \left| < w_i, x > + b_i \right|_+$$

Cybenko, Girosi, ….

Both deep and shallow models can approximate continuous functions, but suffering the curse of dimensionality…

Kolmogorov's Superposition Theorem

Theorem (A. Kolmogorov, 1956; V. Arnold, 1957)

Given $n \in \mathbb{Z}^+$, every $f_0 \in C([0,1]^n)$ can be reprensented as

$$f_0(x_1, x_2, \cdots, x_n) = \sum_{q=1}^{2n+1} g_q \left(\sum_{p=1}^{n} \phi_{pq}(x_p) \right),$$

where $\phi_{pq} \in C[0,1]$ are increasing functions independent of f_0 and $g_q \in C[0,1]$ depend on f_0.

- Can choose g_q to be all the same $g_q \equiv g$ (Lorentz, 1966).
- Can choose ϕ_{pq} to be Hölder or Lipschitz continuous, but not C^1 (Fridman, 1967).
- Can choose $\phi_{pq} = \lambda_p \phi_q$ where $\lambda_1, \cdots, \lambda_n > 0$ and $\sum_p \lambda_p = 1$ (Sprecher, 1972).

If 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.

Curse of Dimensionality

$$y = f(x_1, \ldots, x_d)$$

Curse of dimensionality

Both shallow and deep network can approximate a function of d variables equally well. The number of parameters in both cases depends exponentially on d as $O(\varepsilon^{-d})$.

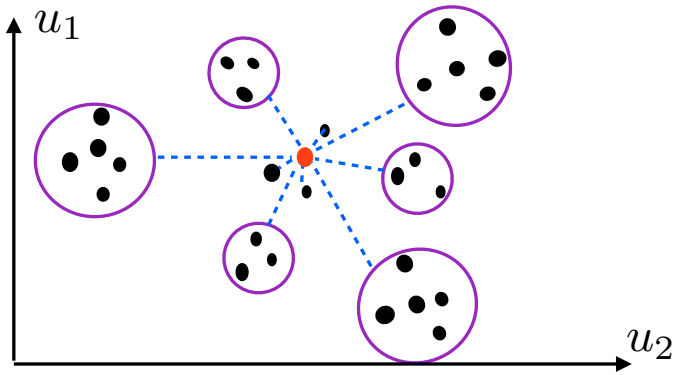
CENTER FOR Brains Minds + Machines

Mhaskar, Poggio, Liao, 2016

A Blessing from Physical world? Multiscale "compositional" sparsity

- Variables $x(u)$ indexed by a low-dimensional u: time/space... pixels in images, particles in physics, words in text...

- Mutliscale interactions of d variables:



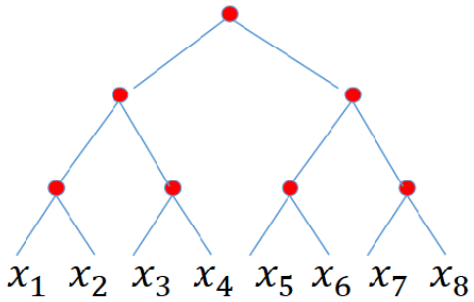
From d^2 interactions to $O(\log^2 d)$ multiscale interactions.

(Or even of constant numbers.)

- Multiscale analysis: wavelets on groups of symmetries. hierarchical architecture.

Hierarchically local compositionality

$$f(x_1, x_2, ..., x_8) = g_3(g_{21}(g_{11}(x_1, x_2), g_{12}(x_3, x_4))g_{22}(g_{11}(x_5, x_6), g_{12}(x_7, x_8)))$$



Theorem (informal statement)

Suppose that a function of d variables is hierarchically, locally, compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$ with the dimension whereas for the deep network dance is $O(d\varepsilon^{-2})$

CENTER FOR Brains Minds + Machines

Mhaskar, Poggio, Liao, 2016

Convolutional Neural Networks (VGG, ResNet etc.) are of this type.

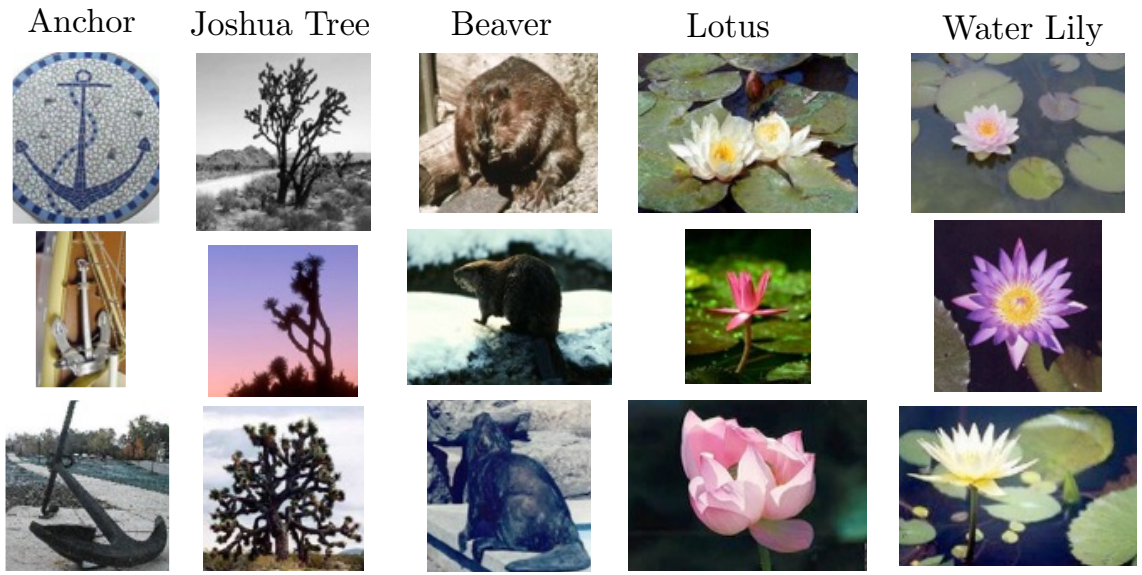
Geometric group invariant properties of deep networks

Contraction within-class level sets toward group invariants when depth grows, while keeping the separation between classes

High Dimensional Natural Image Classification

- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:
- **Classification:** estimate a class label $f(x)$ given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \leq n}$

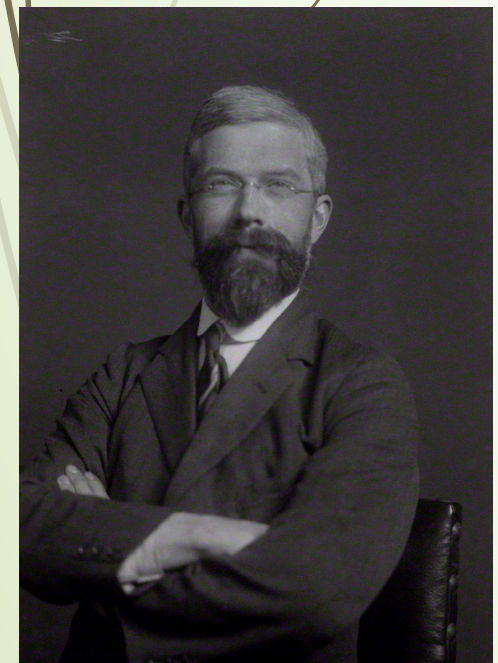
Image Classification $d = 10^6$



Huge variability inside classes

Find invariants

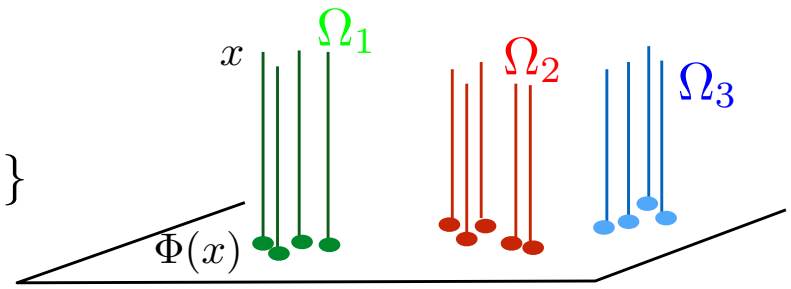
Fisher's Linear Discriminant (1936) (Linear Dimensionality Reduction)



Classes

Level sets of $f(x)$

$$\Omega_t = \{x \ : \ f(x) = t\}$$



If level sets (classes) are parallel to a linear space then variables are eliminated by linear projections: *invariants*.

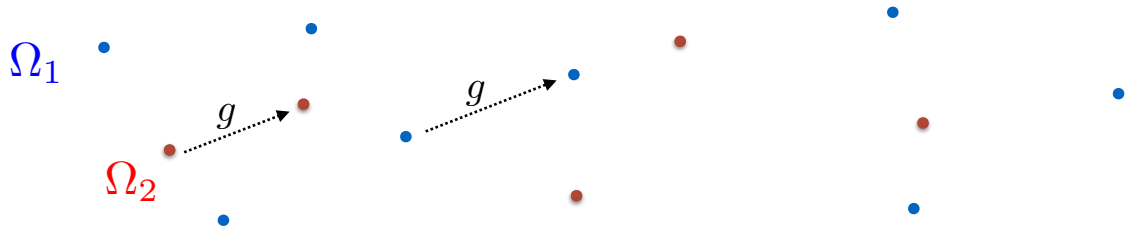
$$\Phi(x) = \alpha \hat{\Sigma}_W^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$$

$$\hat{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \qquad \hat{\Sigma}_W = \sum_k \sum_{i \in C_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Nonlinear Level Set Group Symmetries

Level Set Geometry: Symmetries

- Curse of dimensionality $\Rightarrow$ not local but global geometry
  Level sets: classes, characterised by their global symmetries.



- A symmetry is an operator g which preserves level sets:

$$\forall x \quad , \quad f(g.x) = f(x) \; : \; \text{global}$$

If g_1 and g_2 are symmetries then $g_1.g_2$ is also a symmetry

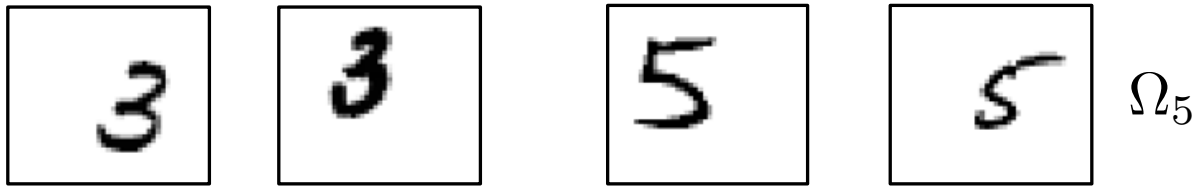
$$f(g_1.g_2.x) = f(g_2.x) = f(x)$$

Level set symmetries lead to groups…

Translation and Deformations

- Digit classification:

$x(u)$ $x'(u) = x(u - \tau(u))$

Ω_3  Ω_5

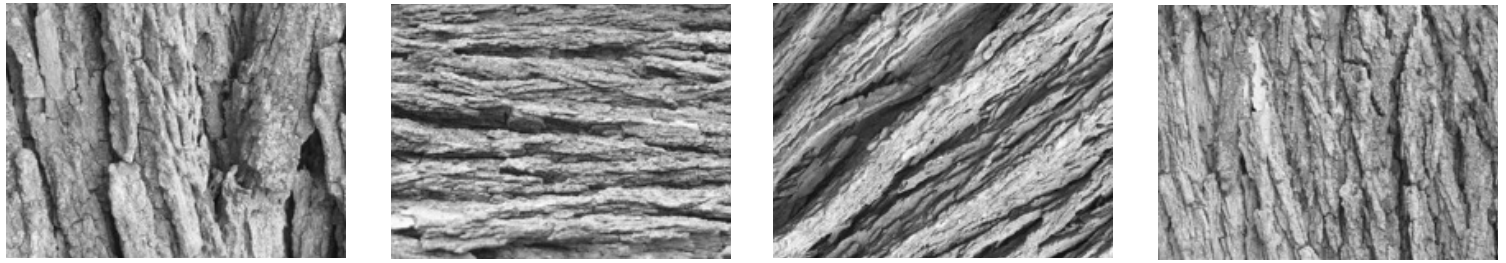
- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Video of Philipp Scott Johnson

https://www.youtube.com/watch?v=nUDIoN-_Hxs

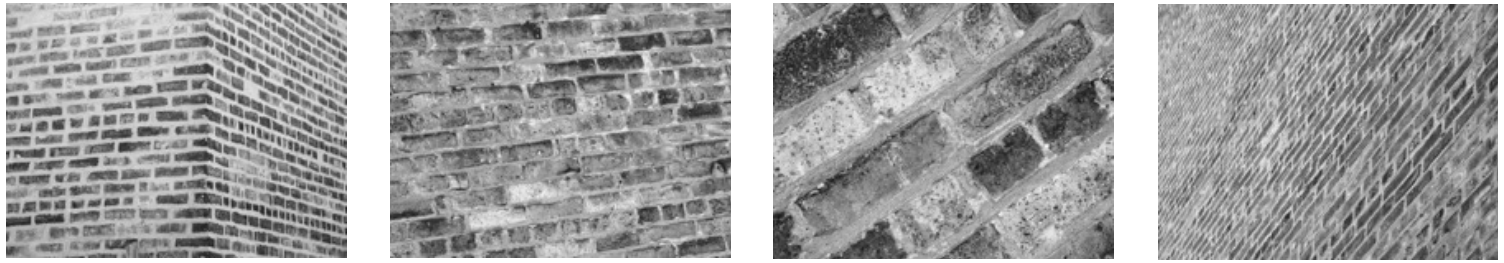
Rotation and Scaling Variability

- Rotation and deformations



Group: $SO(2) \times \mathrm{Diff}(SO(2))$

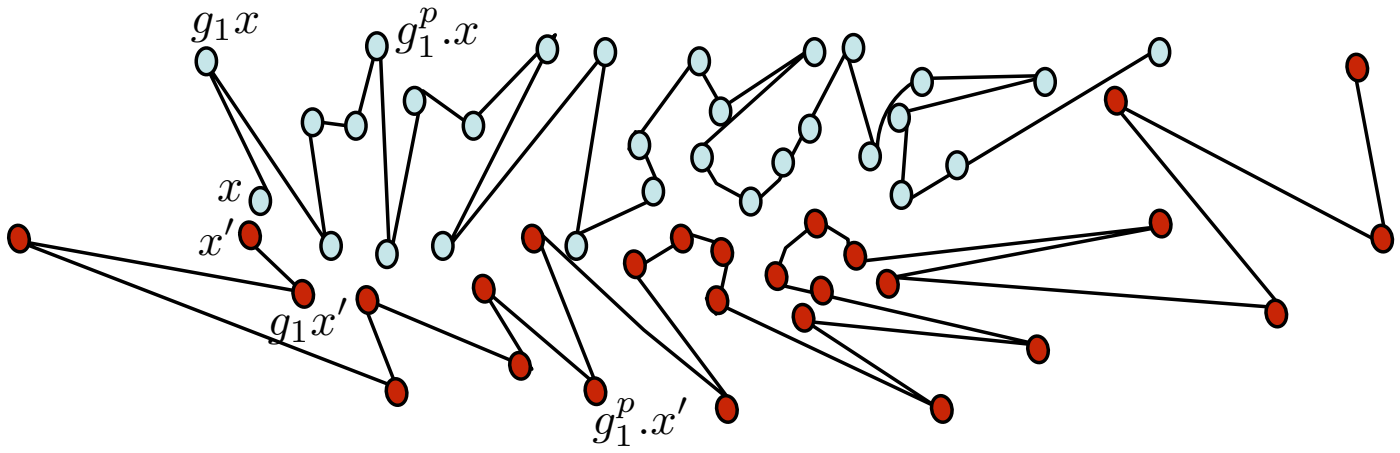
- Scaling and deformations



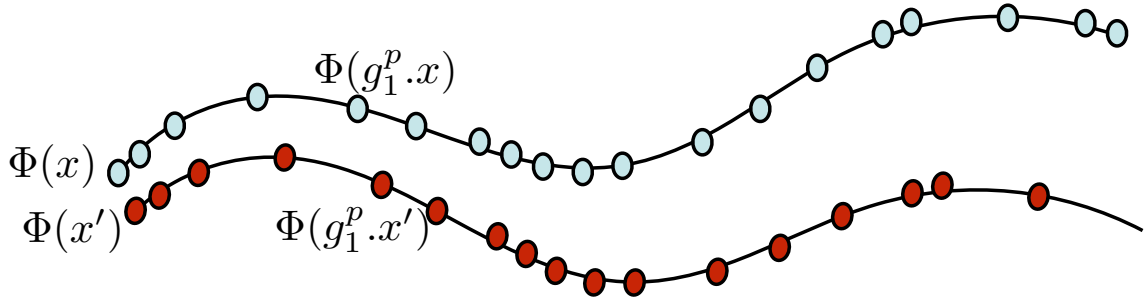
Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

Linearize Symmetries

- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g \in G}$



- Linearise symmetries with a change of variable $\Phi(x)$



- Lipschitz: $\forall x, g \ : \ \|\Phi(x) - \Phi(g.x)\| \leq C \|g\|$

Wavelet Scattering Net

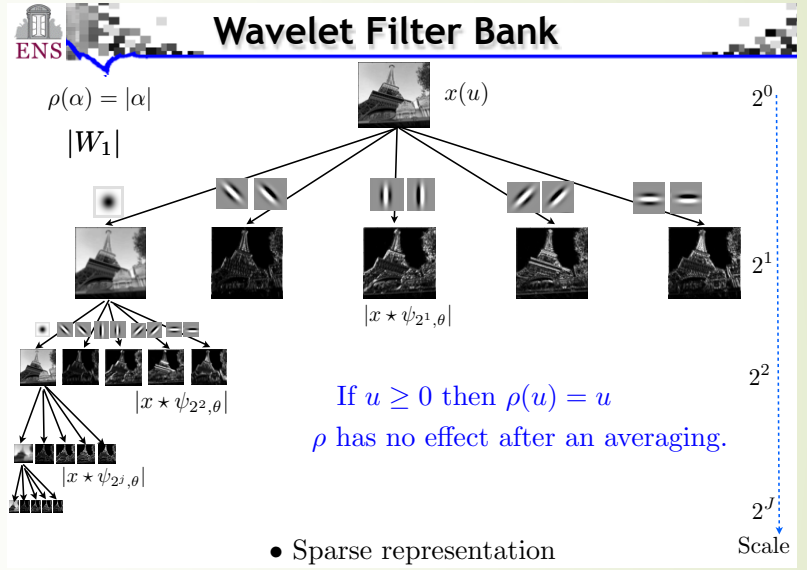
Stephane Mallat et al. 2012

- Architechture:
  - Convolutional filters: band-limited complex wavelets
  - Nonlinear activation: modulus (Lipschitz)
  - Pooling: averaging (L1)
- Properties:
  - A Multiscale Sparse Representation
  - Norm Preservation (Parseval's identity):

$$\|Sx\| = \|x\|$$

  - Contraction:

$$\|Sx - Sy\| \leq \|x - y\|$$



$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \cdots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Invariants/Stability of Scattering Net

- **Translation Invariance** (generalized to **rotation** and **scaling**):

  - The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.

  - Full translation invariance at the limit:

$$\lim_{\phi \to 1} |x \star \psi_{\lambda_1}| \star \phi(t) = \int |x \star \psi_{\lambda_1}(u)| \, du = \|x \star \psi_{\lambda_1}\|_1$$

- Stable Small Deformations:

$$stable\ to\ deformations\ \ x_\tau(t) = x(t - \tau(t))$$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)| \, \|x\|$$

Wiatowski-Bolcskei'15

- Scattering Net by Mallat et al. so far
  - Wavelet Linear filter
  - Nonlinear activation by modulus
  - Average pooling
- Generalization by Wiatowski-Bolcskei'15
  - Filters as frames
  - Lipschitz continuous Nonlinearities
  - General Pooling: Max/Average/Nonlinear, etc.
  - As depth grows, the multiplicative pooling factors leads to full invariances.

Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. Then,

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Summary

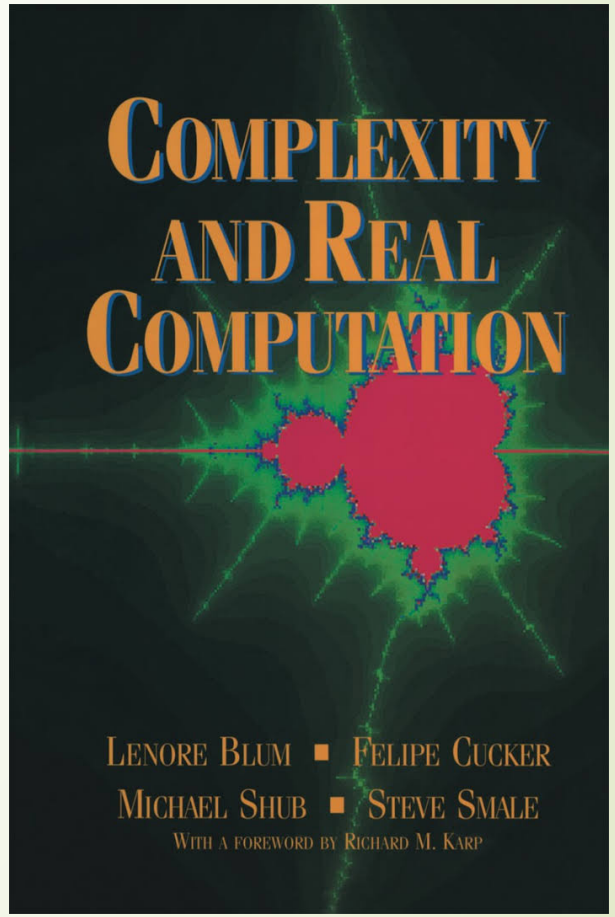
- All these works partially explains the success of CNNs
  - Contraction within level set symmetries toward invariance when depth grows
  - Separation kept between different levels (discriminant)
- Other questions?
  - Can topology be learned with finite information?

Topology can be learned with finite information if the manifold is in good condition

Blum-Shub-Smale models of Real Computation

A Model of Real Computation

- Starting from **Blum, Shub, Smale** (1989)
- It admits inputs and operations (addition, substraction, multiplication, and (in the case of fields) division) of **real (complex) numbers** with *infinite precision*
- "The key importance of the **condition number**, which measures the closeness of a problem instance to the manifold of ill-posed instances, is clearly developed." – Richard Karp



The Condition Number of a Manifold

Throughout our discussion, we associate to $\mathcal{M}$ a condition number $(1/\tau)$ where τ is defined as the largest number having the property: The open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^N$ for every $r < \tau$. Its image Tub_τ is a tubular neighborhood of $\mathcal{M}$ with its canonical projection map

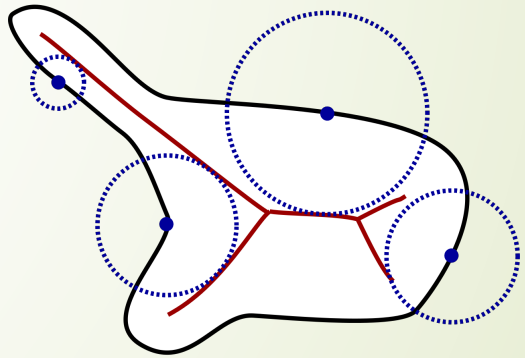
$$\pi_0 : \text{Tub}_\tau \to \mathcal{M}.$$

Smallest Local Feature Size

$G = \{x \in \mathbb{R}^N$ such that $\exists$ distinct $p, q \in \mathcal{M}$ where $d(x, \mathcal{M}) = \|x - p\| = \|x - q\|\}$,

where $d(x, \mathcal{M}) = \inf_{y \in \mathcal{M}} \|x - y\|$ is the distance of x to $\mathcal{M}$. The closure of G is called the medial axis and for any point $p \in \mathcal{M}$ the local feature size $\sigma(p)$ is the distance of p to the medial axis. Then it is easy to check that

$$\tau = \inf_{p \in \mathcal{M}} \sigma(p).$$



Find Homology with Finite Samples

[Niyogi, Smale, Weinberger (2008)]

Theorem 3.1 *Let $\mathcal{M}$ be a compact submanifold of $\mathbb{R}^N$ with condition number τ. Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points drawn in i.i.d. fashion according to the uniform probability measure on $\mathcal{M}$. Let $0 < \epsilon < \tau/2$. Let $U = \bigcup_{x \in \bar{x}} B_\epsilon(x)$ be a correspondingly random open subset of $\mathbb{R}^N$. Then for all*

$$n > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right) \right),$$

the homology of U equals the homology of $\mathcal{M}$ with high confidence (probability $> 1 - \delta$).

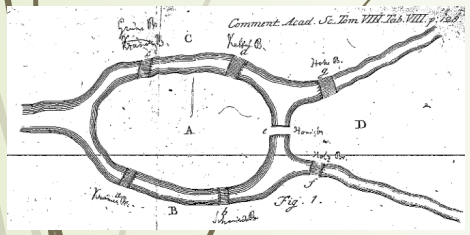
$$\beta_1 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_1))vol(B^k_{\epsilon/4})} \quad \text{and} \quad \beta_2 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_2))vol(B^k_{\epsilon/8})}.$$

Here k is the dimension of the manifold $\mathcal{M}$ and $vol(B^k_\epsilon)$ denotes the k-dimensional volume of the standard k-dimensional ball of radius ϵ. Finally, $\theta_1 = \arcsin(\epsilon/8\tau)$ and $\theta_2 = \arcsin(\epsilon/16\tau)$.

Topological Data Analysis

How to compute topology from finite sample points with noise?

The origin of Topology



- We would like to say that all points within tolerance are the same
- Moreover, all non-zero distances beyond tolerance are the same, i.e. invariant under distortion

- Origins of Topology in Math
  - Leonhard Euler 1736, Seven Bridges of Königsberg
  - Johann Benedict Listing 1847, Vorstudien zur Topologie
  - J.B. Listing (orbituary) Nature 27:316-317, 1883. "qualitative geometry from the ordinary geometry in which quantitative relations chiefly are treated."

Discrete Data Analysis in Metric Space

Fact

We Don't Trust Large Distances!

- In life or social sciences, distance (metric) are constructed using a notion of similarity (proximity), but have no theoretical backing (e.g. distance between faces, gene expression profiles, Jukes-Cantor distance between sequences)
- Small distances still represent similarity (proximity), but long distance comparisons hardly make sense

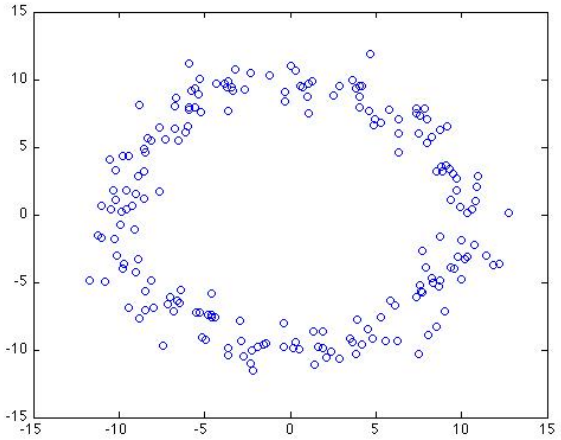
Fact

We Only Trust Small Distances a Bit!

- Both pairs are regarded as similar, but the strength of the similarity as encoded by the distance may not be so significant
- Similar objects lie in neighborhood of each other, which suffices to define topology

Fact

Even Local Connections are Noisy, depending on observer's scale!



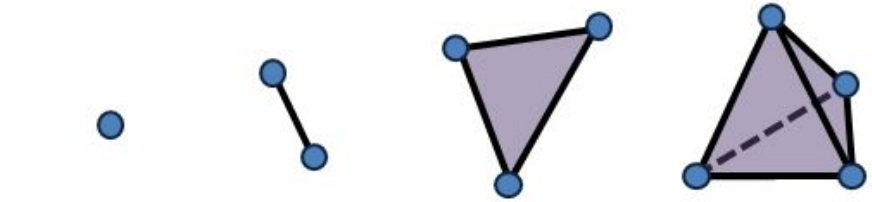
- Is it a circle, dots, or circle of circles?
- To see the circle, we ignore variations in small distance (tolerance for proximity)

Topological Data Analysis

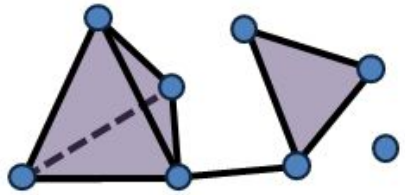
- Distance/metric measurements are noisy
- Physical device like human eyes may ignore differences in proximity
- **Topology** is the crudest way to capture invariants under distortions of distances
- At the presence of noise, one need robust **topology invariant with metric scales:**
  - simplicial complexes for data representation
  - filtration on simplicial complexes
  - persistent homology

Simplicial Complexes

A **simplicial complex** is built from points, edges, triangular faces, etc.

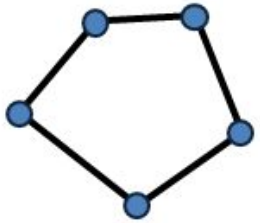


0-simplex 1-simplex 2-simplex 3-simplex (solid)

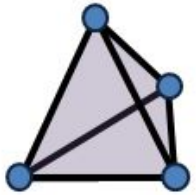


example of a simplicial complex

Homology counts components, holds, voids, etc.



hole



void (contains faces but empty interior)

Homology of a simplicial complex is computable via linear algebra.

Boundary Map and Chain Complexes

C_n =*n-th Chain Group= Formal linear combinations of simplices of the complex*

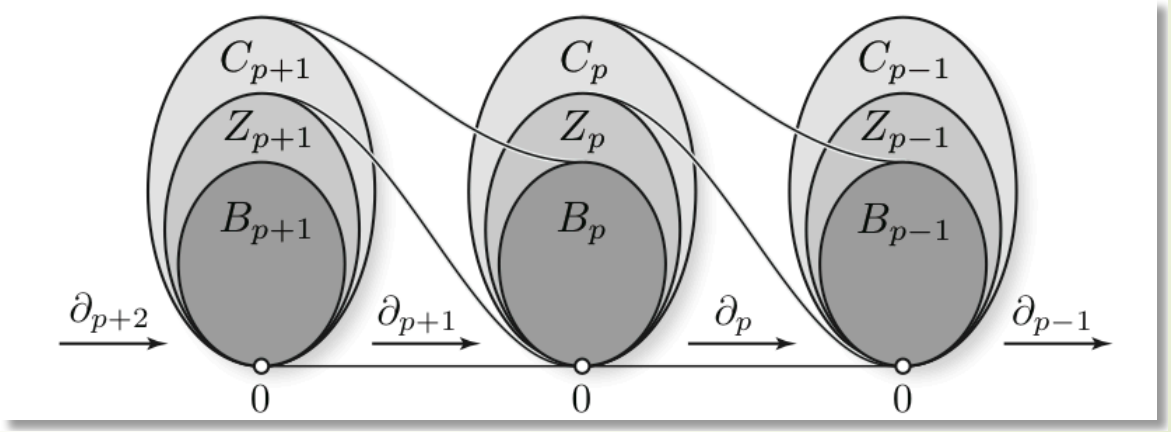
$$\partial_n \sigma_n(\Delta^n) = \sum_{k=0}^{n} (-1)^k [p_0, \cdots, p_{k-1}, p_{k+1}, \cdots p_n]$$

Boundary Map: sends a simplex to a combination of its faces

$$\cdots \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

$$\partial_n \circ \partial_{n+1} = 0_{n+1,n-1}$$

Nilpotency=boundary of boundary is 0



Homology Groups and Betti Numbers

$$H_n(X) := \ker(\partial_n)/\mathrm{im}(\partial_{n+1}) = Z_n(X)/B_n(X),$$

β_k=# of generators of H_k= Betti Number

β_0=#of connected components,

β_1=#of holes,

β_2=# voids, ……

$$\chi = \sum_{k=}(-1)^k \beta_k$$

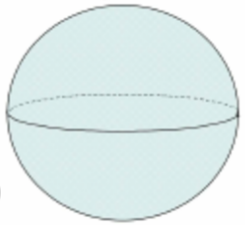
Euler's Characteristic

A solid 2-dimensional blob

β_0=1

$\beta_{i>0}$=0

A sphere



β_0= 1

β_1= 0

β_2= 1

$\beta_{i>2}$= 0

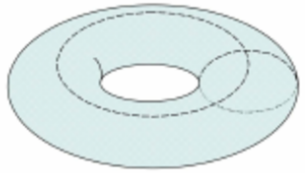
A 2D blob with three holes

β_0= 1

β_1= 3

$\beta_{i>1}$= 0

A torus

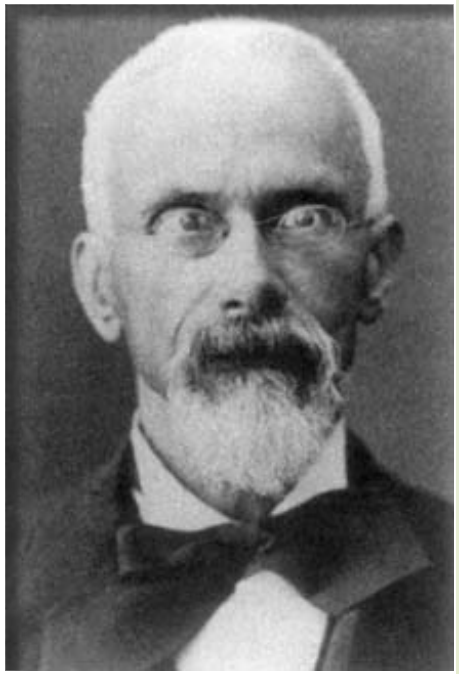


β_0= 1

β_1= 2

β_2= 1

$\beta_{i>2}$= 0



E. Betti, 1823-1892

Hodge Theory for Real (co)-Homology

For inner product spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, consider

$$\mathcal{X} \xrightarrow{A} \mathcal{Y} \xrightarrow{B} \mathcal{Z}.$$

and $\Delta = AA^* + B^*B : \mathcal{Y} \to \mathcal{Y}$ where $(\cdot)^*$ is adjoint operator of $(\cdot)$. If

$$B \circ A = 0,$$

then $\ker(\Delta) = \ker(A^*) \cap \ker(B)$ and *orthogonal* decomposition

$$\mathcal{Y} = \mathrm{im}(A) + \ker(\Delta) + \mathrm{im}(B^*)$$

Note: $\ker(B)/\mathrm{im}(A) \simeq \ker(\Delta)$ is the (real) (co)-homology group ($\mathbb{R} \to$ rings; vector spaces$\to$module).

Abstract Simplicial Complex

Definition (Simplicial Complex)

An abstract simplicial complex is a collection Σ of subsets of V which is closed under inclusion (or deletion), i.e. $\tau \in \Sigma$ and $\sigma \subseteq \tau$, then $\sigma \in \Sigma$.

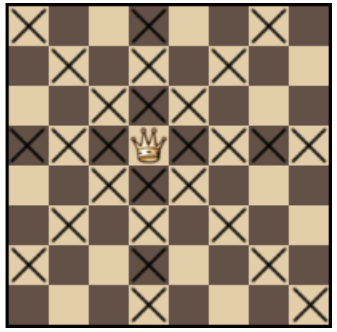
- Chess-board Complex
- Point cloud data:
  - Nerve complex
  - Cech, Rips, Witness complex
  - Mayer-Vietoris Blowup
- Term-document cooccurance complex
- Clique complex in pairwise comparison graphs

Example I

Definition (Chess-board Complex)

Let V be the positions on a Chess board. Σ collects position subsets of V where one can place queens (rooks) without capturing each other.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of "safe" positions, then any subset $\tau \subseteq \sigma$ is also a set of "safe" positions

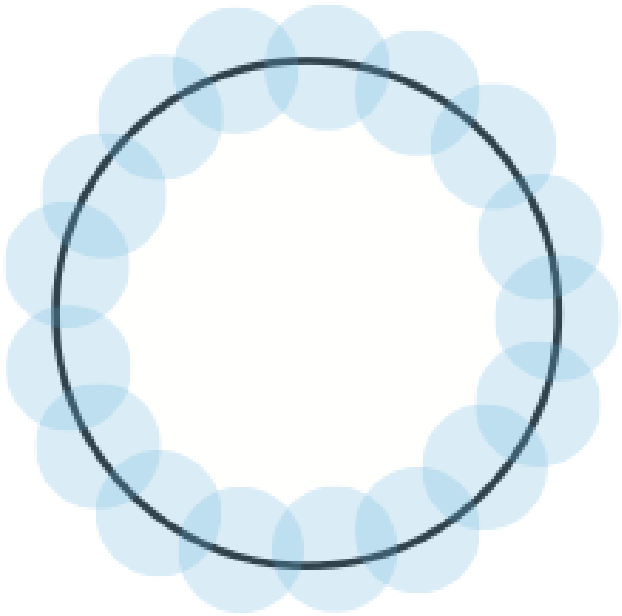


Example II

Definition (Nerve Complex)

Define a cover of X, $X = \cup_\alpha U_\alpha$. $V = \{U_\alpha\}$ and define $\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset\}$.

- Closedness under deletion
- Can be applied to any topological space X
- In a metric space (X, d), if $U_\alpha = B_\epsilon(t_\alpha) := \{x \in X : d(x - t_\alpha) \leq \epsilon\}$, we have Čech complex C_ϵ.
- Nerve Theorem: if every U_I is contractible, then X has the same homotopy type as Σ.



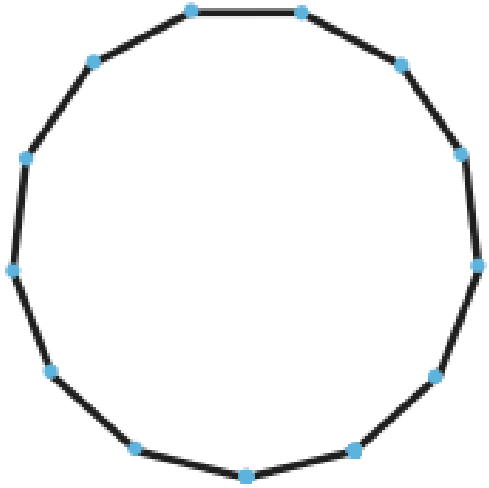


Figure: Čech complex of a circle, C_ϵ, covered by a set of balls.

Example III

- Čech complex is hard to compute, even in Euclidean space
- One can easily compute an upper bound for Čech complex
  - Construct a Čech subcomplex of 1-dimension, i.e. a graph with edges connecting point pairs whose distance is no more than ϵ.
  - Find the clique complex, i.e. maximal complex whose 1-skeleton is the graph above, where every k-clique is regarded as a $k-1$ simplex

Definition (Vietoris-Rips Complex)

Let $V = \{x_\alpha \in X\}$. Define
$VR_\epsilon = \{U_I \subseteq V : d(x_\alpha, x_\beta) \leq \epsilon, \alpha, \beta \in I\}$.

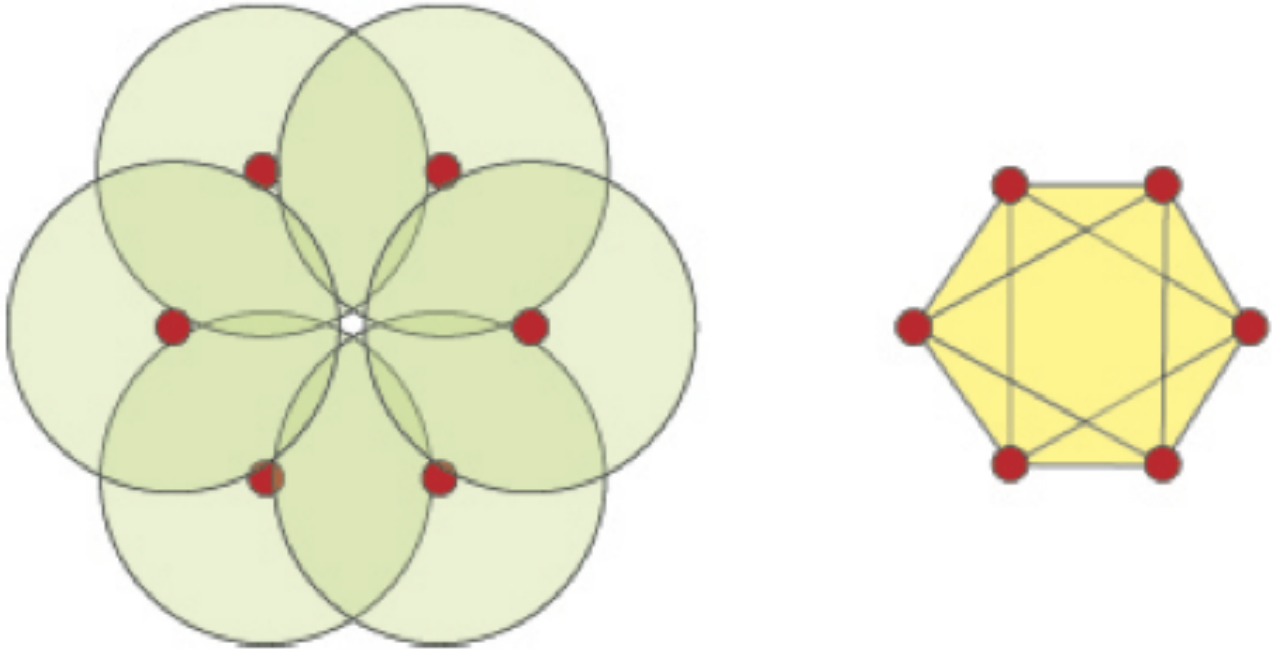


Figure: Left: Čech complex gives a circle; Right: Rips complex gives a sphere S^2.

E.g. Vietoris-Rips Complex of Tolerance Space

Definition (Symmetric Relation Complex)

Let V be a set and a symmetric relation $R = \{(u, v)\} \subseteq V^2$ such that $(u, v) \in R \Rightarrow (v, u) \in R$. Σ collects subsets of V which are in pairwise relations.

- Closedness under deletion: if $\sigma \in \Sigma$ is a set of related items, then any subset $\tau \subseteq \sigma$ is a set of related items
- Generalized Vietoris-Rips complex beyond metric spaces
- E.g. Zeeman's tolerance space
- C.H. Dowker defines simplicial complex for unsymmetric relations

Persistent Homology

- Rips is easier to compute than Cech
  - even so, Rips is exponential to dimension generally
- However Vietoris-Rips CAN NOT preserve the homotopy type as Cech
- But there is still a hope to find a lower bound on homology –

Theorem ("Sandwich")

$$VR_\epsilon \subseteq C_\epsilon \subseteq VR_{2\epsilon}$$

- If a homology group "persists" through $R_\epsilon \to R_{2\epsilon}$, then it must exists in C_ϵ; but not the vice versa.

Betti Numbers and Scales

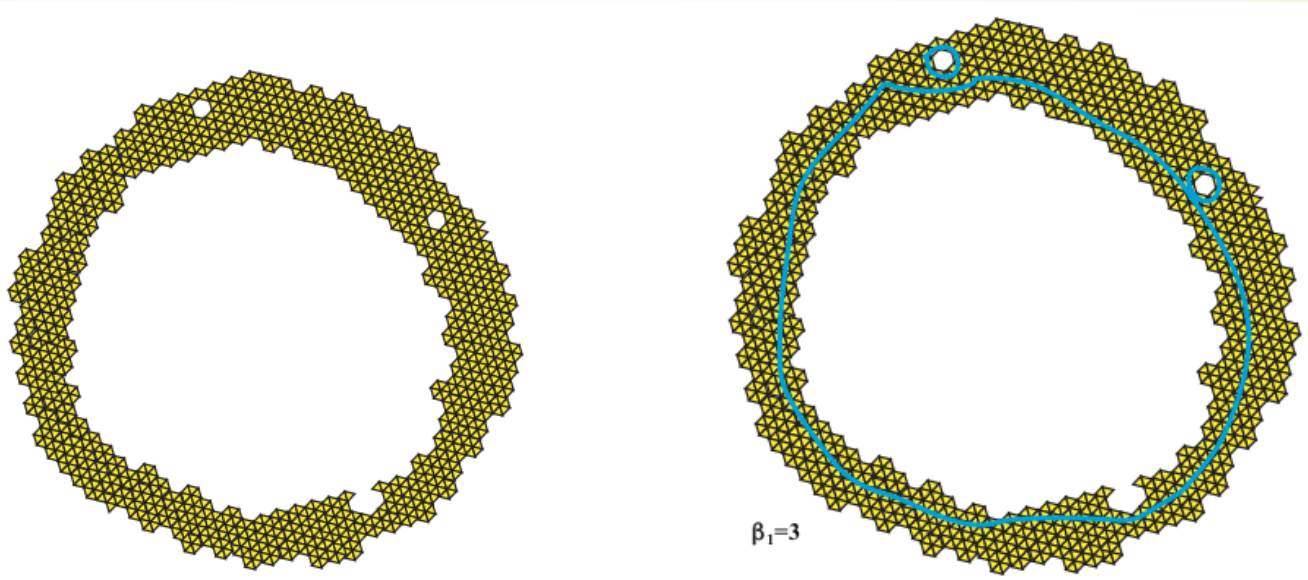


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 3$

Betti Number and Scales

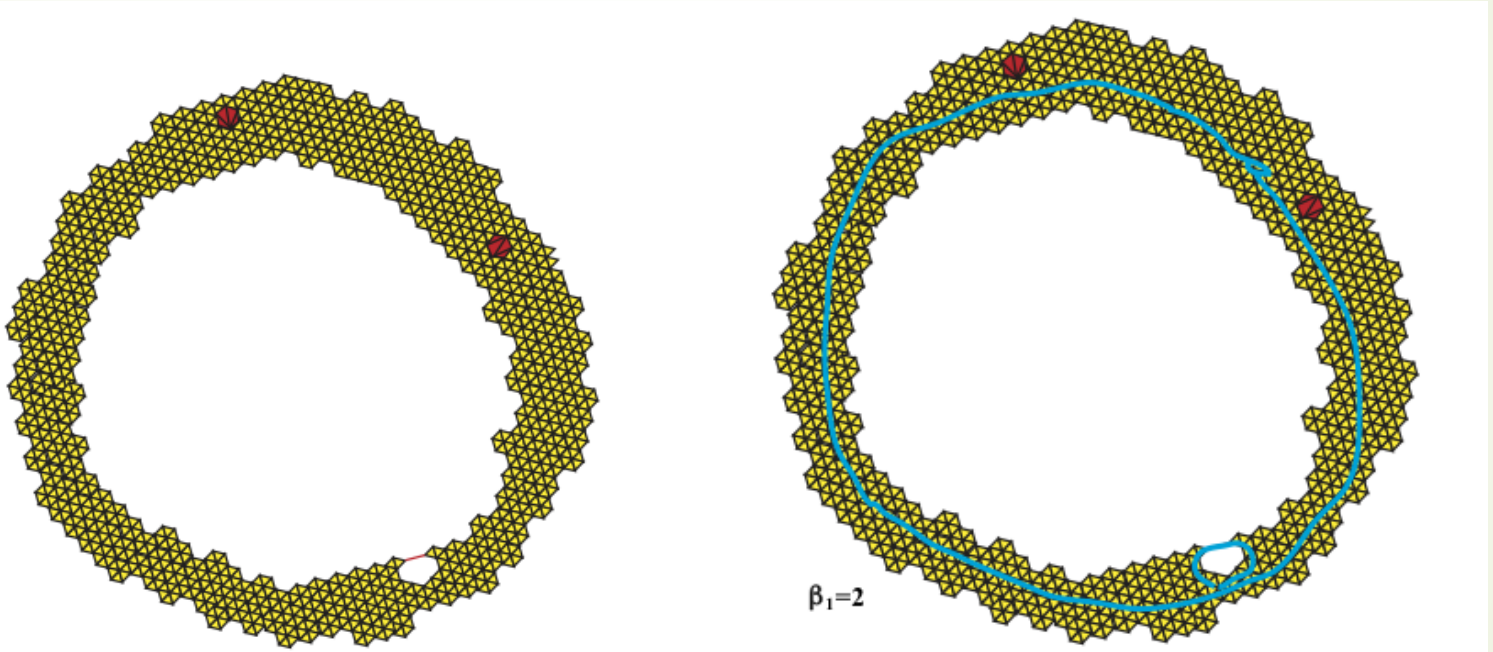
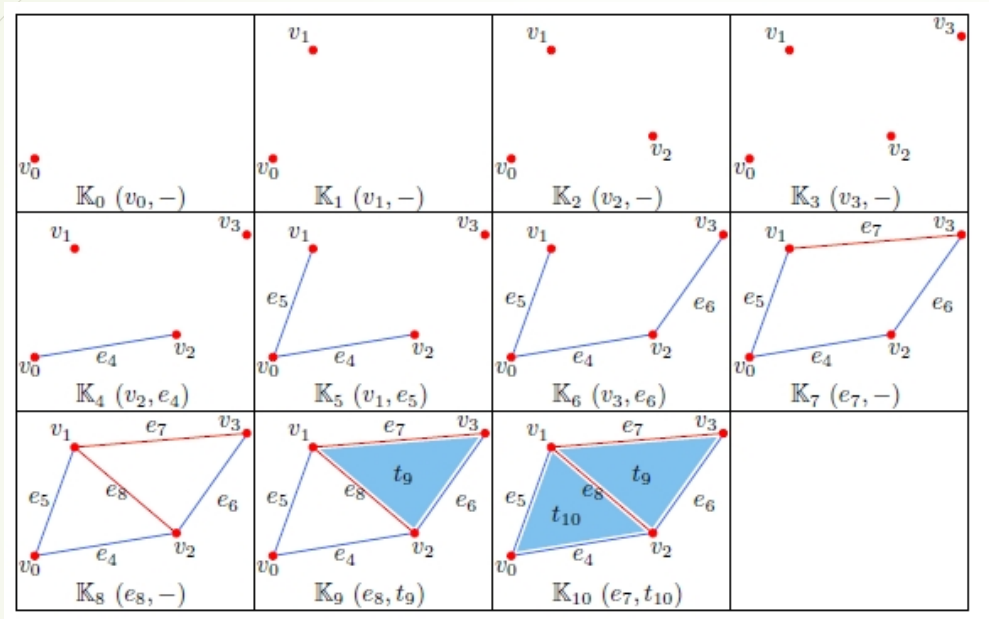
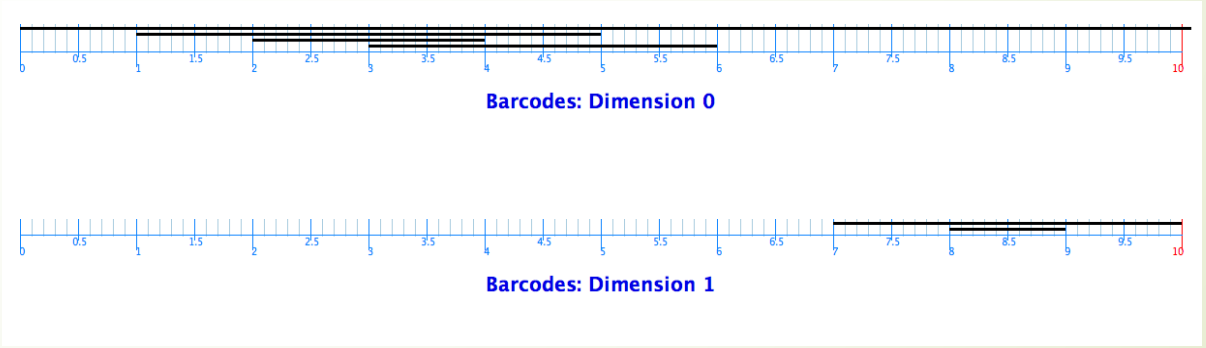


Figure: Scale ϵ_1: $\beta_0 = 1$, $\beta_1 = 2$

Persistent Homology Algorithm





Persistent Homology

- All above gives rise to a filtration of simplicial complex

$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \dots$$

- Functoriality of inclusion: there are homomorphisms between homology groups

$$0 \to H_1 \to H_2 \to \dots$$

- A persistent homology is the image of H_i in H_j with $j > i$.

Curse of Dimensionality and Quantum Algorithms

To construct a Rips-complex of dimension of n points: $O(2^n)$ number of simplices is needed in the worst case => O(poly(n)) in Quantum Algorithms

Quantum Algorithms for Topological Data Analysis

nature COMMUNICATIONS

ARTICLE

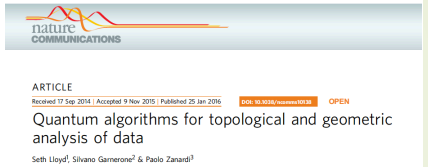
Received 17 Sep 2014 | Accepted 9 Nov 2015 | Published 25 Jan 2016

DOI: 10.1038/ncomms10138 OPEN

Quantum algorithms for topological and geometric analysis of data

Seth Lloyd[1], Silvano Garnerone[2] & Paolo Zanardi[3]

Quantum Algorithm for Persistent Homology: The Sketch


nature COMMUNICATIONS

ARTICLE

Received 17 Sep 2014 | Accepted 9 Nov 2015 | Published 25 Jan 2016 DOI: 10.1038/ncomms10138 OPEN

Quantum algorithms for topological and geometric analysis of data

Seth Lloyd[1], Silvano Garnerone[2] & Paolo Zanardi[3]

0) Store (or compute) distances between data points in a Q-RAM

1) Fix ε, construct a quantum state encoding simplicial complex at the scale ε (Grover Search Algorithm)

2) Find the kernel of the Laplacian to get the Betti Numbers (Quantum Phase estimation Algorithm)

3) Iterate over the ε and look for persistent features across scales

How About computational complexity?!?

Computational Complexity: Classical vs. Quantum

Table 1 | Computational cost comparison.

Procedural steps	Classical cost	Quantum cost
Input pairwise distances, n points	$O(n^2)$ bits	$O(n^2)$ bits
Construct simplicial complex	$O(2^n)$ ops	$O(n^2)$ ops on $O(n)$ qubits
Diagonalize Laplacian/find Betti numbers	$O(2^{2n}\log(1/\delta))$ ops	$O(n^5/\delta)$ quantum ops

δ is the multiplicative accuracy to which the Betti numbers and the eigenvalues of the combinatorial Laplacian are determined. Note the trade-off between the exponential quantum speed-up and accuracy: the quantum algorithms obtain an exponential speed-up over classical algorithms but provide an accuracy that scales polynomially in $1/\delta$ rather than exponentially. This feature arises from the nature of the quantum phase estimation/matrix inversion algorithms, which obtain their exponential speed-up by estimating eigenvectors and eigenvalues using a 'pointer-variable' measurement interaction[38–40]. By contrast, classical algorithms need only keep $O(\log(1/\delta))$ bits of precision, but must perform $O(2^{2n})$ steps to diagonalize $2^n \times 2^n$ sparse matrices.

NATURE COMMUNICATIONS | 6:10138 | DOI: 10.1038/ncomms10138 | www.nature.com/naturecommunications

Our Quantum algorithm provides and exponential speed-up over the classical one!

QUANTUM SUPREMACY……. $|\Psi\rangle$

The Guts of the Quantum Algorithm I


nature COMMUNICATIONS

ARTICLE

Quantum algorithms for topological and geometric analysis of data

Seth Lloyd[1], Silvano Garnerone[2] & Paolo Zanardi[3]

Let s_k a k-simplex we map it onto a quantum state $|s_k\rangle=|j_1,j_2,\ldots,j_n\rangle$ where $j_p=1$ iff p is in s_k

$\mathcal{H}_k^\epsilon$ Space generated by the k-simplex states in the $\boldsymbol{\varepsilon}$-Complex, $|S_k^\epsilon|$ --dimensional

Quantum Pipeline 1: Encoding the $\boldsymbol{\varepsilon}$-Complex

Grover's Search Algorithm:

$$|\psi\rangle_k^\epsilon = \frac{1}{\sqrt{|S_k^\epsilon|}} \sum_{s_k \in S_k^\epsilon} |s_k\rangle,$$

Takes time $O\left(n^2 (\zeta_k^\epsilon)^{-1/2}\right)$ where ζ_k^ϵ is fraction of simplices actually present in the $\boldsymbol{\varepsilon}$-Complex; Classical time $O(2^n)$

The Guts of the Quantum Algorithm II


nature COMMUNICATIONS

ARTICLE

Quantum algorithms for topological and geometric analysis of data

Seth Lloyd, Silvano Garnerone & Paolo Zanardi

Combinatorial Hodge Theory: Betti numbers are the dimensions of the kernels of the $\boldsymbol{\varepsilon}$-complex Laplacian operators (*0*-eigenvectors=Harmonic forms ≅ to Homology classes)

$$\Delta_k = \tilde{\partial}_k^\dagger \tilde{\partial}_k + \tilde{\partial}_{k+1} \tilde{\partial}_{k+1}^\dagger$$

If $B_k^\epsilon = \begin{pmatrix} 0 & \tilde{\partial}_k \\ \tilde{\partial}_k^\dagger & 0 \end{pmatrix}$. then $B^\epsilon = B_1^\epsilon \oplus B_2^\epsilon \oplus \ldots \oplus B_n^\epsilon$. is the $\boldsymbol{\varepsilon}$-complex Dirac operato

$$B^{\epsilon 2} = \Delta_0 \oplus \Delta_1 \oplus \ldots \oplus \Delta_n$$

Quantum Pipeline 2

Run the **Quantum Phase Algorithm** for B^ϵ over the uniform mixture of all simplices Determines the dimensions of $\operatorname{Ker} \Delta_k$. i.e., the Betti's numbers

Classically: $O\left(\binom{n}{k}^2\right) \sim O(2^{2n})$ Quantumly (*n*-sparsity→) $O(n^5)$

Demonstration by 6-photon Quantum Computer [Huang et al. 2018, arXiv:1801.06316]

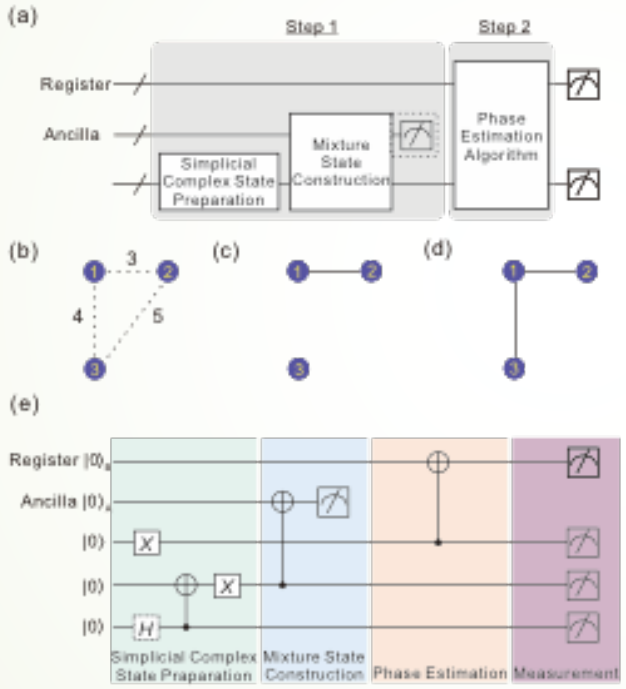


FIG. 2. Quantum circuit for quantum TDA. (a) Outline of the original quantum circuit. (b) A scatterplot including three data points. (c) Graph representation of the 1-simplices state $|\varphi\rangle_1^{\epsilon_1} = |110\rangle$ for $3 < \epsilon_1 < 4$. The first and second data points are connected by an edge. (d) Graph representation of 1-simplices state $|\varphi\rangle_1^{\epsilon_2} = (|110\rangle + |101\rangle)/\sqrt{2}$ for $4 < \epsilon_2 < 5$. The first data point is connected to the second and third points by two edges. (e) Optimized circuit with 5 qubits. The blocks with different colors represent the four basic stages.

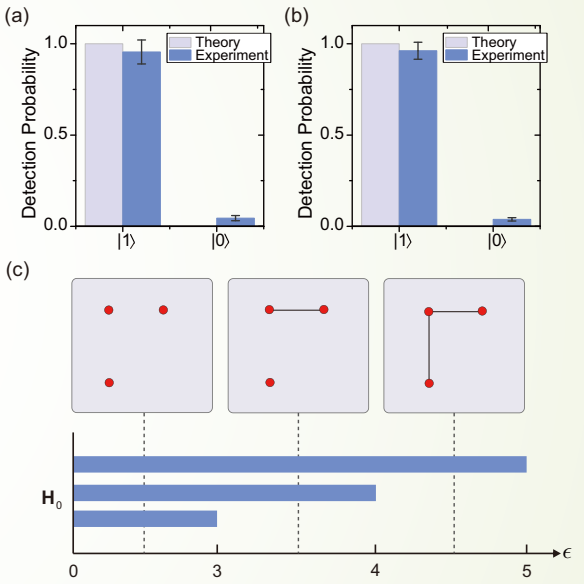


FIG. 4. Final experimental results. The output is determined by measuring the eigenvalue register in the Pauli-Z basis. Measured expectation values (blue bars) and theoretically predicted values (gray bars) are shown for two different 1-simplices state inputs: (a) $|\varphi\rangle_1^{\epsilon_1} = |110\rangle$, (b) $|\varphi\rangle_1^{\epsilon_2} = (|110\rangle + |101\rangle)/\sqrt{2}$. Error bars represent one standard deviation, deduced from propagated Poissonian counting statistics of the raw detection events. (c) The barcode for $0 < \epsilon < 5$. Since no k-dimensional holes for $k \geq 1$ exist at these scales, only the 0-th Betti barcode is given here. For $0 < \epsilon < 3$, there is no connection between each point, so the 0-th Betti number is equal to the number of points. That is, there are three bars at $0 < \epsilon < 3$. At scales of $3 < \epsilon_1 < 4$ and $4 < \epsilon_2 < 5$, the 0-th Betti number are 2 and 1.

Summary

- **Minsky-Papert (1969)** Model:
  - locality/sparsity is a fundamental limit of neural networks
  - XOR is not possible for single layer perceptrons
  - Topological computation needs non-local (global) information
- Deep Neural Networks
  - Both shallow and deep networks are universal, but suffer the curse-of-dimensionality
  - Locality/sparsity may avoid the curse-of-dimensionality (**Poggio** et al. 2017)
  - Geometric group invariants grow with depth (**Mallat** et al. 2012; **Bolcskei** et al. 2017)
- **Blum-Shub-Smale (1989)** Model of Real Computation
  - Condition number of a manifold is a fundamental limit of topology learning with finite samples
- Topological Data Analysis
  - Persistent homology suffers the curse-of-dimensionality
- Quantum algorithms for topological data analysis
  - A polynomial complexity algorithm for homology computation (**Lloyd** et al. 2016)
  - A 3-point demonstration by 6-photon quantum computer (**Huang** et al. 2018)

Thank you!