On studying neural network expressiveness using topological data analysis and knot theory

Alexandre Louvet

December 1, 2020

Abstract

In this paper we summarize the state of the art on the question of neural network expressiveness both on the theoretical approach to the problem with the study of universal approximators and some practical approaches using topological data analysis and trajectories. We then propose an analysis of the question from a knot theory perspective and share results using studied methods for datasets in dimension 3 and 4.

Contents

1 Neural network expressiveness **4**
1.1 Definition 4
1.2 Universal Approximator 5
1.2.1 Sigmoidal functions 5
1.2.2 ReLU functions 7

2 Homology **10**
2.1 The idea of homology 10
2.2 Δ-complexes 12
2.3 Simplicial homology 14
2.4 Homology in our research 16

3 Topological data Analysis **19**
3.1 Introduction to TDA method's 19
3.1.1 Density Clustering 19
3.1.2 Low dimensional subsets 20
3.2 Persistent homology 22

4 Knot theory **25**
4.1 Fundamental concepts 25
4.2 Knot determinant 25
4.2.1 Definition 25
4.2.2 Algorithms 25

5 Measuring neural network expressiveness **27**
5.1 Using topological data analysis 27
5.2 Using trajectories 27

6 The study of trajectories from a knot theory perspective **29**
6.1 Methodology 29
6.2 Algorithms 29
6.3 Results 29

7 Extending the study of expressiveness with topological data analysis **31**
7.1 Methodology 31
7.2 Algorithms 31
7.3 Results 31

List of Figures

1 Some images of the MNIST dataset 16
2 A plot of the swiss roll dataset. We have $d = 3$ but the support of the data S is of dimension $r = 2$ 21
3 On this plot we have a dataset of dimension 2 but there is a support of dimension 1 that concentrates a big part of the mass 21

1 Neural network expressiveness

1.1 Definition

Let I_n denote the n-dimensional unit cube $[0,1]^n$ and $\mathcal{F}(I_n, \mathbb{R})$ be the space of functions from I_n to $\mathbb{R}$. We want to study the density of the subsets S_f of $\mathcal{F}(I_n, \mathbb{R})$ that can be written as follows:

$$S_f = \{G_N(x) \in \mathcal{F}(I_n, \mathbb{R}) \mid G(x) = \sum_{i=1}^{N} \alpha_i f(y_j^T x + \theta_j)\},\ N \in \mathbb{N}$$

depending on the choice of $f \in \mathcal{F}(\mathbb{R}, \mathbb{R})$. In the previous equation $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$, y^T is the transpose of y and $y^T x$ is the inner product of y and x.

The study of neural network expressiveness consists of the problem described above when f is a function used as an activation function for neural network. The study of density can be on the whole set $\mathcal{F}(I_n, \mathbb{R})$ or on subsets of it such as $\mathbb{C}(I_n, \mathbb{R})$ the set of continuous functions from I_n to $\mathbb{R}$.

In particular if S_f is dense in a subset $A \subseteq \mathcal{F}(I_n, \mathbb{R})$ we will say that a single-layer feed-forward neural network (Fig. 1) with f as its activation function is a *universal approximator* of A. Considering a neural network has a finite number of nodes neural network expressiveness also consists of the study of the rate of approach of the approximation, i.e. the sudy of

$$\lim_{N \to \infty} H(N) = \max_{h \in A}(\min_{G_n \in S_f}(\| G_n - h \|)) \text{ with } \| \cdot \| \text{ the canonical norm on } \mathcal{F}(I_n, \mathbb{R})$$

The study of that limit and especially of its asympatotic approximation gives an idea of the efficiency of the approximator, i.e. the amount of node to add to the network to improve the approximation.

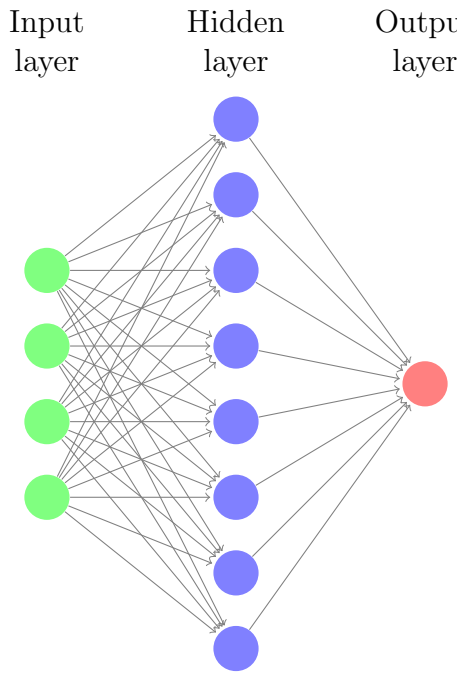


Fig 1: A single-layer feed-forward neural network with $n = 4$ and $N = 8$

1.2 Universal Approximator

In this section we will study the different subsets on which the logistic and ReLU functions acts as universal approximators.

1.2.1 Sigmoidal functions

We say that a function $\sigma \in \mathcal{F}(I_n, \mathbb{R})$ is a sigmoidal function if:

$$\sigma(x) \to \begin{cases} 0 & \text{as } t \to +\infty \\ 1 & \text{as } t \to -\infty \end{cases}$$

The sigmoidal functions include the logistic function defined as:

$$f(x) = \frac{1}{1+e^{-x}}$$

widely used as an activation function for neural networks.

The first study of neural network expressiveness with sigmoidal functions date back to by G.Cybenko in 1989 [1]. He proves that S_σ for σ a sigmoidal function is dense in regards of the supremum norm in $C(I_n, \mathbb{R})$. The demonstration goes as follows.

We denote $M(I_n)$ the space of signed regular Borel measures on I_n

Definition 1 *σ is discriminatory if $\mu \in M(I_n)$ and*

$$\forall y \in \mathbb{R}^n,\ \theta \in \mathbb{R} \int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0 \implies \mu = 0$$

Theorem 1 *Let σ be a contininuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{i=1}^{N} \alpha_i \sigma(y^T x + \theta_i)$$

are dense in $C(I_n, \mathbb{R})$

PROOF: Let $S \subset C(I_n)$ be the set of the function of the form $G(x)$. S in a linear subset of $C(I_n)$. Let us show that the closure of S is $C(I_n)$.
Assume it is not the case. Then the closure of S, denoted R, is a proper subspace of $C(I_n)$. Using the Hahn-Banach theorem, there esists L a bounded linear functional on $C(I_n)$ with $L \neq 0$ and $L(R) = L(S) = 0$
Using the Riesz Representation Theorem, we obtain:

$$L(h) = \int_{I_n} h(x) d\mu(x)$$

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. Since $\sigma(y^T x + \theta_i) \in R$, we have

$$\forall y, \theta \int_{I_n} \sigma(y^T x + \theta) d\mu(x) = 0$$

Since σ is discriminatory, we have $\mu = 0$ and $L = 0$ follows! Hence the closure of S is $C(I_n)$ and by definition S is dense in $C(I_n)$ □

Now let us show that sigmoidal functions are discriminatory.

Lemma 1 *Any bounded, measurable sigmoidal function, a, is discriminatory. In particular, any continuous sigmoidal function is discriminatory.*

PROOF: First note $\forall x, y, \theta, \phi$

$$\sigma_\lambda(x) = \sigma(\lambda(y^T x + \theta) + \phi) \begin{cases} \to 1 & \text{for } y^T x + \theta > 0 \text{ as } \lambda \to +\infty \\ \to 0 & \text{for } y^T x + \theta < 0 \text{ as } \lambda \to +\infty \\ = \sigma(\phi) & \text{for } y^T x + \theta = 0 \end{cases}$$

Thus $\sigma_\lambda(x)$ converges pointwise and boundedly to:

$$\gamma(x) \begin{cases} = 1 & \text{for } y^T x + \theta > 0 \\ = 0 & \text{for } y^T x + \theta < 0 \\ = \sigma(\phi) & \text{for } y^T x + \theta = 0 \end{cases}$$

as $\lambda \to +\infty$

Let $\Pi_{y,\theta} = \{x \mid y^T x + \theta = 0\}$ and let $H_{y,\theta} = \{x \mid y^T x + \theta > 0\}$. Lesbegue bounded convergence theorem gives us:

$$\begin{aligned} 0 &= \int_{I_n} \sigma_\lambda(x) d\mu(x) \\ &= \int_{I_n} \gamma(x) d\mu(x) \\ \sigma(\phi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \end{aligned}$$

for all ϕ, θ, y

Fix y, we write

$$F(h) = \int_{I_n} h(y^T x) d\mu(x)$$

Note that F is a bounded function on $L^\infty(\mathbb{R})$ since μ is a signed mesure. By chosing h as the indicator function on $[\theta, \infty[$, we have:

$$F(h) = \int_{I_n} h(y^T x) d\mu(x) = \mu(\Pi_{y,-\theta}) + \mu(H_{y,-\theta}) = 0$$

By linearity $F(h) = 0$ for indicator function on any interval and hence for any simple function (sum of indicator functions) and since simple functions are dense in $L^\infty(\mathbb{R}), F = 0$. In particular it is true for the bounded function $s(u) = sin(m.u)$ and $c(u) = cos(m.u)$. It gives:

$$F(s + ic) = \int_{I_n} cos(m^T x) + isin(m^T x) d\mu(x) = \int_{I_n} exp(im^T x) d\mu(x) = 0$$

for all m. Therefore the fourier transform of μ is 0 and μ must be 0. Hence σ is discriminatory. □

This proves that any function of $C(I_n, \mathbb{R})$ can be approximated by a single-layer network with sigmoidal functions as activation function.

In 1991, Hornik extended in [2] gave the proof for bounded non-constant functions of $F(I_n, \mathbb{R})$. The proof is very similar to the original one.

1.2.2 ReLU functions

ReLU functions stands for Rectified Linear Units, there are formed of two pieces of linear functions. They gained interest recently by showing convincing results in a lot of different applications.

In 2016 Arora et al. [3] showed a version of the universal approximation theorem for piecewise linear function that includes ReLU functions.

Definition 2 *We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous piecewise linear (PWL) if there exists a finite set of polyhedra whose union is $\mathbb{R}^n$, and f is affine linear over each polyhedron (implies continuity because affine regions are closed and cover $\mathbb{R}^n$)*

Proposition 1 *Every function in $L^q(\mathbb{R}^n), (1 \leq q \leq \infty)$ can be arbitrarily approximated in the L^q norm (which for a function f is given by $||f||_q = (\int |f|^q)^{1/q}$) by a ReLU Deep Neural Network.*

Proof:
We know that any ReLU DNN represents a PWL function, let's prove the converse.

By theorem 1 in [4] any piecewise linear function $f : \mathbb{R}^n \to \mathbb{R}$, can be written:

$$f = \sum_{j=1}^{p} s_j(\max_{i \in S_j} l_i)$$

with l_i $(i \leq i \leq k)$ linear functions, S_i $(1 \leq i \leq p) \subseteq \{1, ..., k\}$ with $\forall i, |S_i| \leq n + 1$ and $\forall j \in \{1, ..., p\}, s_j \in \{-1, +1\}$

It means that any PWL convex function can be represented as a linear combination of at most $n + 1$ affine pieces. That's to say a ReLU DNN with size $n + 1$.

Let $p \in \mathbb{N}^*$, let $f \in L^p(\mathbb{R}^n)$, consider the function sequence:

$$f_n(x) = (x - \tfrac{k}{n})f(\tfrac{k}{n}) + (1 - x + \tfrac{k}{n})f(\tfrac{k+1}{n}) \text{ with } \tfrac{k}{n} \leq x < \tfrac{k+1}{n}, n \geq 1$$

$f_n \xrightarrow[n\to\infty]{} f$, and $\forall n$, f_n is a PWL continous function. Therefore the continuous PWL functions are dense in $L^p(\mathbb{R}^n)$.

Since any PWL function $\mathbb{R}^n \to \mathbb{R}$ is representable by a ReLU DNN, unniversal approximation theorem follows from the fact that the family of continuous PWL functions is dense in any $L^p(\mathbb{R}^n)$ space for $1 \leq p \leq \infty$.

□

In this article, Arora et al. also give a higher bound for the depth of the network. they show that the network required for any function $f \in L^q(\mathbb{R}^n)$ is at most $\lceil log_2(n + 1) \rceil$.

In 2018, Hannin & Selke [5] showed that the minimal width a ReLUneural network must have in order to approximate any function $f : \mathbb{R}^n \to \mathbb{R}^m$ is $n + m$. Giving a lower bound to the width of the ReLU network.

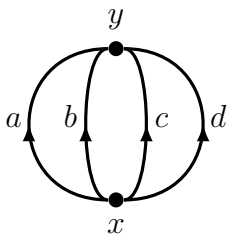
2 Homology

The study of homology is part of the field of algebraic topology. We will first introduce the concept of homology, then take a look at the theoretical concept behind it by introducing Δ-complex: a primitive topolgical structure that allow the study of homology. Then with the introduction of simplicial homology we will complete the presentation of the fundamental objects used in our research and in the study of homology. We will then reconsider these tools in the context of our research. This part heavily relies on the book *Algebraic Topolgy* by A. Hatcher [6] and the class given by P. Albin at University of Illinois [7] with the same book as class material.

2.1 The idea of homology

In algebraic topology, the most fundamental object to study is the homotopy, that roughly studies "paths". This study leads to the definition of the fundamental group $\pi_1(X)$ that gives information on the topological stucture of a space X for low-dimensional space. However the fundamental group $\pi_1(X)$ can not for instance distinguish the different spheres $\mathcal{S}^n, n \geq 2$. A workaround for that problem would be to generalize the notion of homotopy to higher dimensions with different groups $\pi^n(X)$ generalizing the concept of homotopy in higher dimension.

However the homotopy groups suffers a serious drawback in their difficult computationality for high dimensions. Even for simple spaces like spheres, the calculation of $\pi_i(\mathcal{S}^n)$ turns out to be very difficult for $i > n$. That is why homology exists as a computable alternative homotopy.



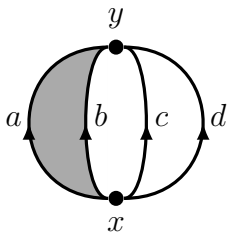
Consider the space X_1 in the figure drawn above. We can see two points x and y as well as four edges a, b, c and d. Consider the loops formed by travelling along the edges, for instance the path ab^{-1} is a loop with x as the basepoint. Something to consider is that the loop $b^{-1}a$ is basically the same loop but starting from a different basepoint, y in that case. By abelianizing, we can consider cycles instead of loops without any basepoint.

The abelian groups having only one operation, we will then switch to additive notations to remain in accordance with Hatcher's notations. With that new notation we can write equalities such as $(a - c) + (b - d) = (a - d) + (b - c)$. This is justified by the fact that from an algebraic point of view, there is no difference between these

two cycles.

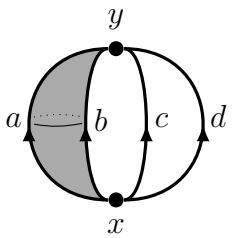
Let us call any linear combination of edges, chains. Then the condition for a chain $ka + lb + mc + nd$ with $k, l, m, n \in \mathbb{N}$ to be a cycle is $k + l + m + n = 0$.

Now let us start to formalize these concepts. Consider C_0 the free abelian group spanned by the vertices x, y and let C_1 be the free abelian group spanned by the edges a, b, c, d. One can define an homomorphism $\delta : C_1 \to C_0$ that sends any basis element of C_1 to $x-y$. Thus we have $\delta(ka+lb+mc+nd) = (k+l+m+n)a-(k+l+m+n)y$ and $ker(\delta) = \{$ cycles of X_1 $\}$. $\{a-b, b-c, c-d\}$ forms a basis for this kernel, i.e. any cycle is a linear combination of the most obvious cycles of X_1, that conveys the information that X_1 has three visibles "holes" formed by the space between its four edges.



Now fill the hole between edges a and b with a 2-cell, namely A to create a new space X_2. Considering it as being oriented clockwise, its boundary is $a - b$. The cycle $a-b$ is homotopic to a point since it can now be contracted by sliding over A, it no longer encloses a hole in G. That suggests that homology should consider only the quotient group of what we found previously by $a-b$. In this quotient group the cycles $a - c$ and $b - c$ are equivalent. Which is consistent with the fact that they are homotopic in X_2.

Algebraically we can now consider a pair of homomorphism δ_1, δ_2 such that $C_2 \xrightarrow{\delta_2} C_1 \xrightarrow{\delta_1} C_0$ where C_2 is the cyclic group spanned by A and $\delta_2(A) = a-b$. The quotient group we are interrested in is $ker(\delta_1)/Im(\delta_2)$, i.e. the 1-dimensional cycles modulo those that are boundaries. We will call it the homology group $H_1(X_2)$. I can also be computed in X_1 by considering C_2 to be zero since there are no two cells in X_2. $H_1(X_2)$ now only admits two generators $b - c$ and cd expressing the geometric fact that the number of holes reduced from 3 to 2.



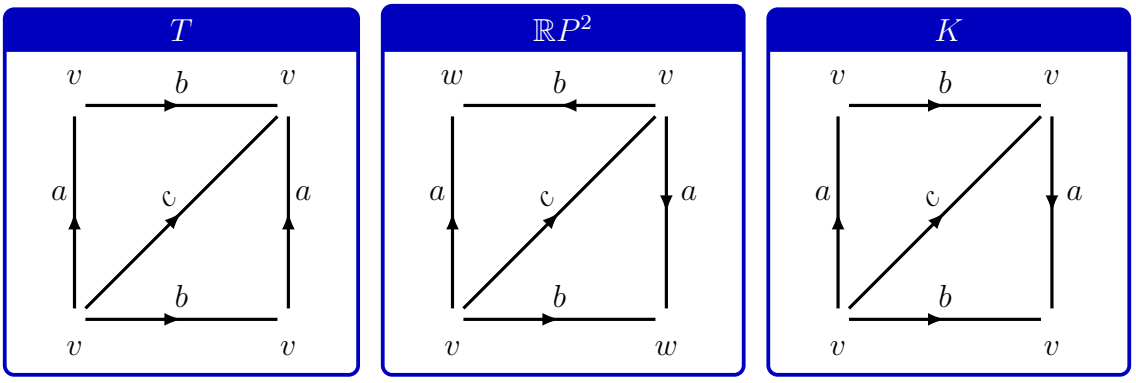
Suppose we attach another 2-cell between a and b to create X_3, namely B. C_2 now consists of linear combinations of A and B and $\delta_2(A) = \delta_2(B) = a - b$. In the one hand, $H_1(X_3) = H_1(X_2) \approx \mathbb{Z} \times \mathbb{Z}$, but this time $ker(\delta_2) \neq 0$ and $A - B$ is its generator, hence we have $H_2(X_3) = ker(\delta_2) \approx \mathbb{Z}$. Topologically the cycle $A - B$ is equivalent to a sphere and it detects the presence of a "hole" enclosed by this sphere rather than a circle.

The pattern we can see appear with these examples is rather clear. The n-cell complexes of X forms free abelian groups $C_n(X)$ and one can define homomorphisms $\delta_n : C_n \to C_{n-1}$ to define homology groups $H_n(X) = ker(\delta_n)/Im(\delta_{n+1})$. The only problem now is to define δ_n for any n. If it is simple for small n (head minus tail for a vertex), it becomes rather complicated when n grows and more complex polyhedral cells appear in X. The most efficient approach is to decompose polyhedra into simplices which allows simple orientation and boundary computation. For that purpose we will define nice cellular complexes that will be the fundamental element for homology computation in the homology theories presented after.

2.2 Δ-complexes

The most important theory of algebraic topology is called simplicial homology, since it is technically complicated we will first introduce a simpler version of it called simplicial homology. The natural definition spaces of simplicial homology is called Δ-complexes that we will introduce in this part.

The projective plane, the klein bottle and the torus can be obtained from squares by identifying edges and giving them orientation as draw below.



In fact any closed surface can be constructed this way, i.e. by cutting a polygon along its diagonals and identifying pairs of edges. The idea between Δ-complexes is to generalize this idea to any dimension. The n-dimensional analog of the triangle is called the n-simplex.This is the smallest convex set in a Euclidean space $\mathbb{R}^m$ containing $n + 1$ points $v_0, ..., v_n$ that do not lie in a hyperplane of dimension less than n. The points v_i are the vertices of the simplex and the simplex is denoted $[v_0, ..., v_n]$. The standard n-simplex is:

$$\Delta^n = \{(t_0, ...t_n) \in \mathbb{R}^{n+1} \mid \sum_i t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i\}$$

Its vertices are the unit vectors along the coordinate axes.

It is important to keep track of an ordering for the vertices and an n-simplex will always be considered with an ordering for its vertices. This will allow us to give an orientation for its edges $[v_i v_j]$ with respect to the ordering of the indices. It also determines a canonical linear homeomorphism from the standart n-simplex Δ^n into any other n-simplex preserving the order of the vertices: $(t_0, ..., t_n) \mapsto \sum_i t_i v_i$. The corfficients are the barycentric coordinates of the points $sum_i t_i v_i$ in $[v_0, ..., v_n]$.

If we delete one of the vertices of an n-simplex $[v_0, ..., v_n]$, the remaining n vertices span an (n-1)-simplex, called face of $[v_0, ..., v_n]$ that will conventionnaly keep the same orientation as in $[v_0, ..., v_n]$. The union of all the face of Δ^n is the boundary of Δ^n, noted $\delta\Delta^n$. The open simplex $\mathring{\Delta}^n$ is $\Delta^n - \delta\Delta^n$, is the interior of Δ^n.

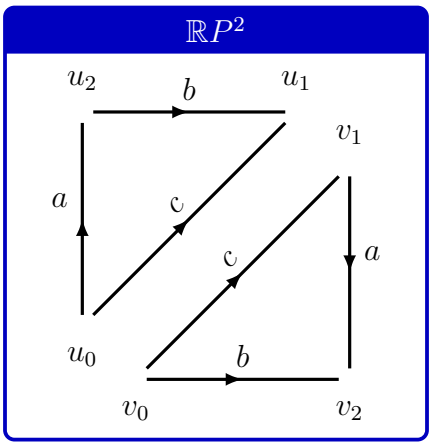
A Δ-complex on a topological space X is a collection of maps $\sigma_\alpha : \Delta^n \to X$ with the following properties:

(i) The restriction $\sigma_\alpha \mid \mathring{\Delta}^n$ is injective, and each point of X is in the image of exactly one such restriction $\sigma_\alpha \mid \mathring{\Delta}^n$.

(ii) Each restriction if σ_α to a face of Δ^n is one if the maps $\sigma_\beta : \Delta^{n-1} \to X$ (identifying the face of Δ^n with the face of Δ^{n-1} using the canonical linear homeomorphism between them and preserving the vertices ordering).

(iii) A set $A \subset X$ is open if and only if $\sigma_\alpha^{-1}(A)$ is open in Δ^n for each σ_α.

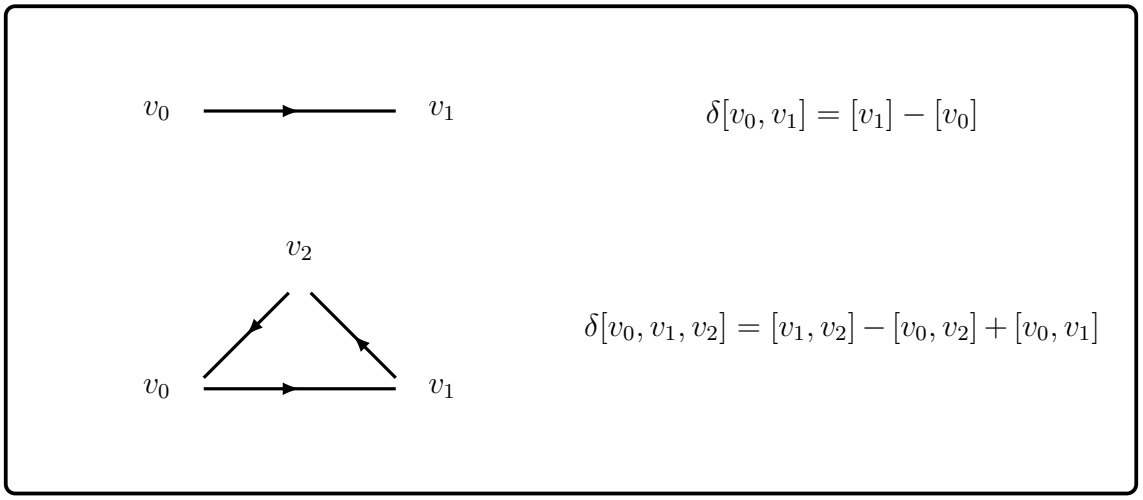
The earlier decompositions of the torus, projective plane, and Klein bottle into two triangles, three edges, and one or two vertices allow well-definess of associated Δ-complex structures with a total of six σ_α's for the torus and Klein bottle and seven for the projective plane. The orientation of edges are compatible with an ordering of the vertices (example below). This ordering determines the maps σ_α.



2.3 Simplicial homology

Let us now define the simplicial homology groups of a Δ-complex X. Let $\Delta_n(X)$ be the free abelian group with basis the open n-simplices e^n_α of X. The elements of $\Delta_n(X)$ are called n-chains and are of the form $\sum_\alpha n_\alpha e^n_\alpha$ with $n_\alpha \in \mathbb{Z}$. Equivalently, this could be written as $\sum_\alpha n_\alpha \sigma_\alpha$ where $\sigma_\alpha : \Delta^n \to X$ is the characteristic map of e^n_α with image the closure of e^n_α. Such a sum can be considered as a finite collection, or "chain", of n-simplices in X.

The boundary of the n-simplex $[v_0, ..., v_n]$ consists of the various $(n-1)$-dimensional simplices $[v_0, ..., \hat{v_i}, ..., v_n]$ (indicates all $v_k, 1 \leq k \leq n$ but i). We might wish that the boundary of $[v_0, ..., v_n]$ be the $(n-1)$-chain formed by the sum of the faces $[v_0, ..., \hat{v_i}, ..., v_n]$. It turns out to be betters to insert sign to take into account orientations. It allow to coherently orient all the faces of the simplex. That is why we let the boundary be $\sum_i (-1)^i [v_0, ..., \hat{v_i}, ..., v_n]$.



We define a boundary homomorphism for a general Δ-complex X, $\delta_n : \Delta_n(X) \to \Delta_{n-1}(X)$ by specifying its values on the absis elements:

$$\delta_n(\sigma_\alpha) = \textstyle\sum_i (-1)^i \sigma_\alpha \mid [v_0, ..., \hat{v_i}, ..., v_n]$$

The image set of δ_n is $\Delta_{n-1}(X)$ since each restriction $\sigma_\alpha \mid [v_0, ..., \hat{v_i}, ..., v_n]$ is the caracteristic map of an $(n-1)$-simplex of X.

Lemma 2 *The composition* $\Delta_n(X) \xrightarrow{\delta_n} \Delta_{n-1}(X) \xrightarrow{\delta_{n-1}} \Delta_{n-2}(X)$ *is zero.*

Proof:
We have $\delta_n(\sigma) = \sum_i (-1)^i \sigma \mid [v_0, ..., \hat{v_i}, ..., v_n]$, hence

$$\begin{aligned}\delta_{n-1}\delta_n(\sigma) = &\sum_{j<i} (-1)^i (-1)^j \sigma \mid [v_0, ..., \hat{v_j}, ..., \hat{v_j}, ..., v_n] \\ &+ \sum_{j>i} (-1)^i (-1)^{j-1} \sigma \mid [v_0, ..., \hat{v_i}, ..., \hat{v_j}, ..., v_n] = 0.\end{aligned}$$

The last equality comes from the switch of i and j in the second sum.

□

We end up with a sequence of homomorphisms of abelian groups:

$$... \to C_{n+1} \xrightarrow{\delta_{n+1}} C_n \xrightarrow{\delta_n} C_{n-1} \to ... \to C_1 \xrightarrow{\delta_1} C_0 \xrightarrow{\delta_0} 0$$

with $\delta_n \delta_{n+1} = 0$ for each n (cf Lemma 2).We call such a sequence a chain complex. The equation $\delta_n \delta_{n+1} = 0$ is equivalent to the inclusion Im $\delta_{n+1} \subset$ Ker δ_n. We can define the n^{th} homology group of the chain complex to be the quotient group $H_n =$ Ker δ_n/Im δ_{n+1}. The elements of Ker δ_n are called cycle and elements of Im δ_{n+1} are called boundaries.Elements of H_n are cosets of Im δ_{n+1} called homology classes. Two cycles representing the same homology class are said to be homologous, it means their difference is a boundary.

In the case of $C_n = \Delta_n(X)$, the homology group Ker δ_n/Imδ_{n+1} will be denoted $H_n^\Delta(X)$ and called the n^{th} simplicial homology group of X.

Example: Let us take $X = T$, the torus with the Δ-complex structure pictured in section **2.2**. For recall, the structure has one vertex v, three edges a, b, c and two 2-simplices U and V. $\delta_1 = v - v = 0$ so $H_0^\Delta(X) \approx \mathbb{Z}$. Since $\delta_2 U = a + b - c = \delta_2 L$ and $\{a, b, a+b-c\}$ is a basis for $\Delta_1(T)$, $H_1^\Delta(X) = \frac{\Delta_1(X)}{\delta_2(U)}$, it follows that $H_1^\Delta(X) \approx \mathbb{Z} \oplus \mathbb{Z}$ with basis the homology classes $[a]$ and $[b]$. Since there are no 3-simplices, $H_2^\Delta(T)$ is equal to Kerδ_2, which is infinite cyclic generated by $U - L$ since $\delta(pU + qL) = (p+q)(a+b-c) = 0$ iff $p = -q$.
Therefore:

$$H_n^\Delta(X) \approx \begin{cases} \mathbb{Z} \oplus \mathbb{Z} & \text{for } n = 1 \\ \mathbb{Z} & \text{for } n = 0, 2 \\ 0 & \text{for } n \geq 3 \end{cases}$$

Proposition 2 *For $n \in \mathbb{N}^*$, $H_n^\Delta(\mathcal{S}^n) \approx \mathbb{Z}$ with $\mathcal{S}^n$ the n-dimension sphere.*

Proof: A Δ-complex structure can be otained on $\mathcal{S}^n$ by taking two copies of Δ^n and identifying boundaries using the identity map. Using previous notations, we have Kerδ_n an infinite cycle generated by $U - L$. Therefore $H_n^\Delta(\mathcal{S}^n) \approx \mathbb{Z}$. □

The simplicial homology is defined for simplical complexes. These are the Δ-complexed whose simplices are uniquely determined by their vertices. Each n-simplex has $n+1$ distinct vertices and no other n-simplex has the same set of vertices. A simplicial complex can be described as a set X_0 of vertices and sets X_n of n-simplices which are $(n+1)$-elements subsets of X_0. It is required that each $(k+1)$-element subset of the vertices of an n-simplex in X_n is a k-simplix in X_k.

Compared with Δ-complexes, simplicial complexes require more computation. For example a simplicial complex structure for the torus would require 14 triangles, 21 edges and 7 vertices. It can be shown that any Δ-complex can be subdivided in a simplical complex.

It is worth mentionning that simplicial homology has flaws such as the one mentionend before. That is why in the current theory of homology, simplicial homology has been replaced with a more complete theory of homologies named singular homology. The improvements of simplicial homology brought by singular are not useful for our research, that is why we will not be discussing this theory in this paper. The theory is however thoroughly explained in [6].

2.4 Homology in our research

Our research is concerned with the capacity of neural wetwork to properly express after training different difficulty of problems depending on their width. There would be different approach to chosing a way to measure this difficulty that would relate to different applied problems. The first approach would be to look at the dimension of the problem.

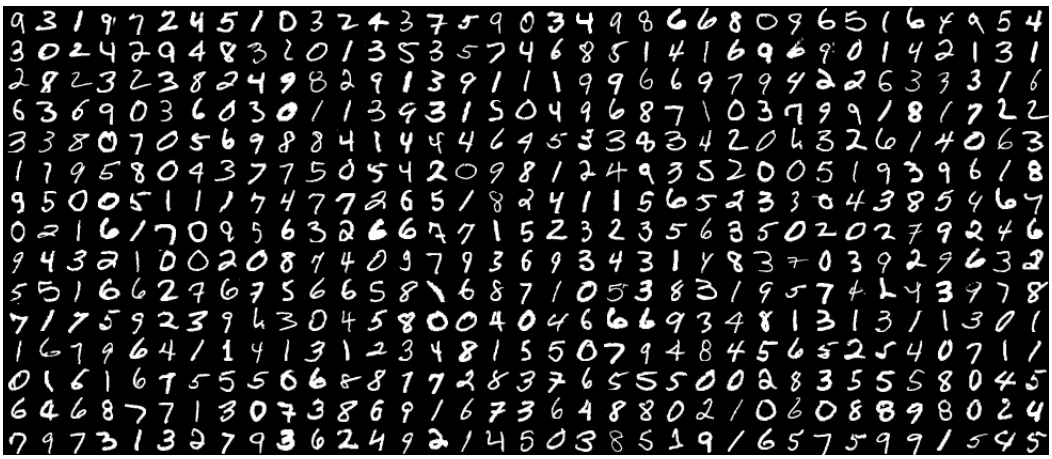


Figure 1: Some images of the MNIST dataset

For instance the MNIST problem, aims to classify 28x28 pixels images of handwritten digits in 10 classes representing the 10 digits (see Figure 1). For instance in 2003 in [8], a team of researcher used a simple neural network of 2 layers with respectively 100 and 10 hidden units to classify the MNIST dataset (they used convolutional layers before inputing data in their network making the input dimension even greater) to obtain an accuracy of 98.81%. One could expect that solving the same problem of image classification for instance and using different image size and depth of networks would be an interresting way to study the problem of expressiveness.

However two major problems occur in that case. How to choose the classification problem with the certainty that the problem itself is not going to be a problem simpler than it looks. This relates to an hypothesis called the manifold hypothesis. The manifold hypothesis holds roughly as follows: "real-world high-dimensional data lie on low-dimensional manifolds embedded within the high-dimensional space". Meaning that complex data that needs a lot of parameters to be perfectly described are caracterized by few simple parameters up to continuous deformations defined as above. Considering this hypothesis that has not been disproved, a certain image recognition problem such as MNIST could be relying on the separation of a space of smaller dimension that the 784 dimensions space that you would expect from 28x28 pixels images.

The second problem is the problem of thresholding for accuracy since it is almost impossible to reach 100%. One would have to chose a value over which one considers the problem as correctly classified and below which it is considered not classified. That choice would be rather arbitrary and it would be hard to evaluate the value of the results as far as they might change radically with a slight change in the choice of the threshold.

This point is non-specific to the choice of accuracy as a measure since in any case the training of a neural network is a process with a random component involved and independently of the measure choice, the problem of training artifacts and imprecicions might rise. Although one should try to chose a measure that will reduce as much as possible this incontrollable component in the research.

For that purpose, we have chosen to use homology of the data as our measure. Using homology garantees an independence from the problem chosen since homologies can take different forms du to its topological nature. Moreover if the manifold hypothesis happends to be true, the study of homology in rather small dimension and how neural networks of different depth perform on it might have applications even in high dimension problems. Unlike accuracy the perfect classification in terms of homology is something reachable since homology is studied without regard to continuous deformations ($\approx$ topologically), errors and imprecision will mostly be ignored.

3 Topological data Analysis

Topolgical Data Analysis, also called TDA, is an approach to the analysis of datasets using techniques from topology. The idea behind it is to provide with a framework to rigorously study the "shape" of a dataset. The main tool is persistent homology, that allows the computation of homology of a point cloud dataset. This tool will be presented in detail after a general presentation of the field of TDA. This chapter is based on the paper of L. Wasserman [9].

3.1 Introduction to TDA method's

Topological Data Analysis definition is somewhat vague and if it is very clear that persistent homology belong to this field, the exact set of methods, theories and result that belongs to this field is quite imprecise. According to this imprecision we chose to give the following definition to TDA: It's a "large class of data analysis method that uses the notion of shape and connectivity" [9]. In his paper Wasserman details different approach to this question of shape and conenctivity study starting with density clusters that we will only briefly cover since these methods are of no use for the purpose of our research. We will then in this section present the ideas and tools behind the study of low-dimension support for high dimension sets and their study with TDA's tools. That second part is here in support of the manifold hypothesis discussed briefly in section **2.4**.

3.1.1 Density Clustering

In his paper, Wasserman covers three types of clustering. More clustering methods could be considered as part of TDA but for explanation purposes of what TDA is, the firt two methods will be sufficient as they will introduce paradigms we will find in persistent homology.

The first method introduced is level set clustering. The idea behind this name is the study of the density of random samples regarding a threshold. Formally consider $X_1, ..., X_n$ a random sample of a distribution P with density p where $X_i \in \mathcal{X} \subset \mathbb{R}^d$. For $t \geq 0$ we define:

$$L_t = \{x : p(x) > t\}$$

The density clusters at level t, denoted C_t, are the connected components of L_t. Once can define an estimated L_t, denoted $\hat{L}_t$ as follows:

$$\widehat{L}_t = \{x : \widehat{p}(x) > t\}$$

where $\widehat{p}(x)$ is an estimator for the density p, such as the kernel density:

$$\widehat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K(\frac{||x - X_i||}{h})$$

where $h > 0$ is the bandwidth and K the kernel.

To find the clusters, one need to find the connected components of $\widehat{L}_t$. Let $I_t = \{i : \widehat{p_h}(X_i) > t\}$. Create a graph with vertices $(X_i : i \in I_t)$. Now put an edge between two vertices X_i and X_j if $||X_i - X_j|| \leq \varepsilon$ where $\varepsilon > 0$ is a parameter ($\varepsilon = 2h$ works well in practice).

It appears the choice of t is crucial in the level set clusters method and there are different possibilities to chose it. First one can chose some prescribed fraction $1 - \beta$ of the total mass. Another idea is to look at clusters at all levels t. This will lead us to the idea of density trees.

The idea of density trees is to create a concept to study level set clustering for all t. That concept will give us both a representation of the structure of the data as well as a way to compare its topological similarity with another set.

The set of all density clusters $\mathcal{C}$ has a tree structure, indeed if $A, B \in \mathcal{C}$ then either $A \subset B$ or $B \subset A$ or $A \cap B = \emptyset$. The tree shows the number of level sets at some level t. The number of branches of the tree correponds to the number of connected components. Two density trees have the same "shape" if their tree structure is the same. Given a tree T_p the distance on the tree is defined as:

$$d_{T_p}(x, y) = |p(x) + p(y) - 2m_p(x, y)|$$

where

$$m_p(x, y) = sup\{t : \exists C \in C_i \mid x, y \in C\}$$

is called the merge height. For any two clusters $C_1, C_2 \in T_p$, we define $\lambda_1 = \{t : C_1 \in C_t\}$ and λ_2 analogously. We then define the tree distance function on T_p as:

$$d_{T_p}(C_1, C_2) = \lambda_1 + \lambda_2 - 2m_p(C_1, C_2)$$

where

$$m_p(C_1, C_2) = sup\{\lambda \in \mathbb{R} : \exists C \in T_p \mid C_1, C_2 \subset C\}$$

Since d_{T_p} defines a distance on the tree, it induces a topology on T_p. Given two densities p and q we can say that T_p is equivalent to T_q, denoted $T_p \approx T_q$ if there exists an homeomorphism between the two trees.

3.1.2 Low dimensional subsets

The point of study of low dimensional subset relates to the manifold hypothesis and more generally the study of the support of the distribution P when it is only a set S of dimension r with $r < d$ (see Figure 2). Sometimes the support of P is of dimension exactly d but there exists a set S of dimension $r < d$ which has a high concentration of mass (see Figure 3).

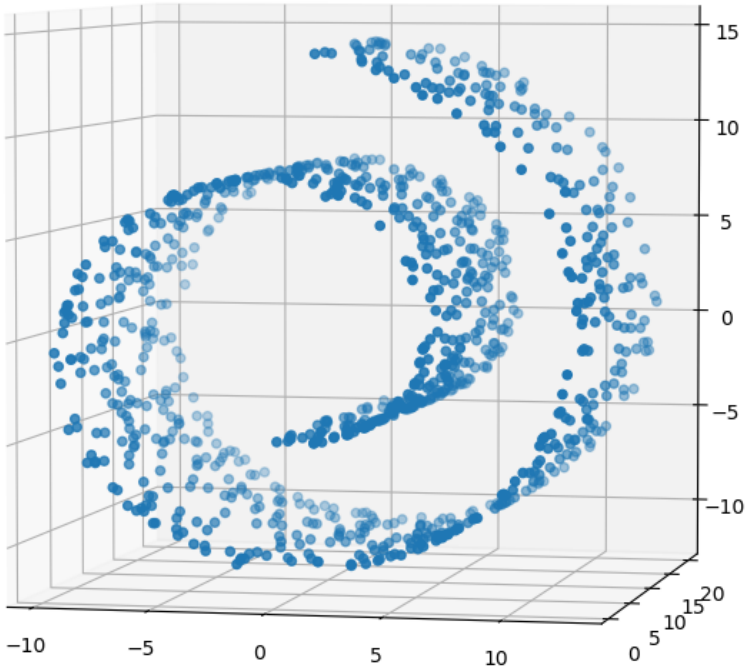


Figure 2: A plot of the swiss roll dataset. We have $d = 3$ but the support of the data S is of dimension $r = 2$

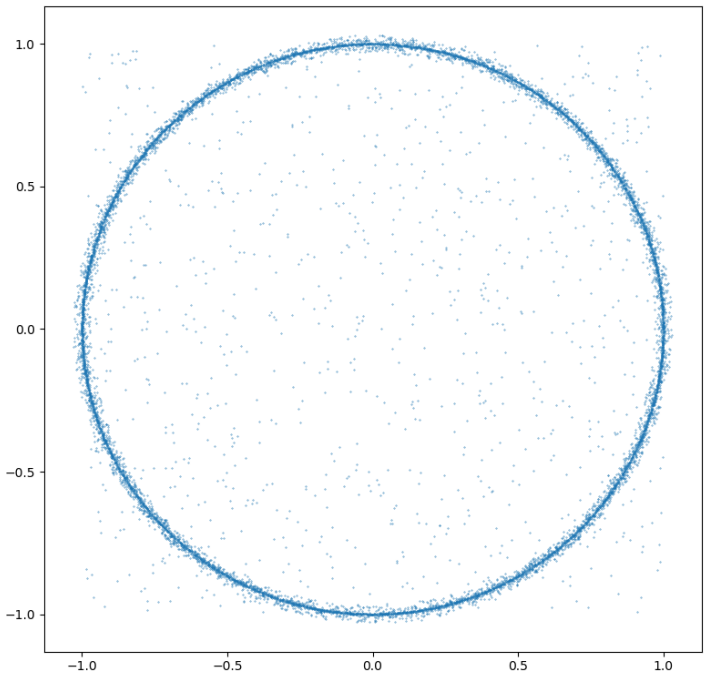


Figure 3: On this plot we have a dataset of dimension 2 but there is a support of dimension 1 that concentrates a big part of the mass

An estimator of S is $\widehat{S} = \bigcup_{i=1}^{n} B(X_i, \varepsilon_n)$. The estimator $\widehat{S}$ is d-dimensional but converges to S. The Hausdorff distance $H(A, B)$ between two sets A and B is:

$$H(A, B) = \inf\{\varepsilon : A \subset B \oplus \varepsilon \text{ and } B \subset A \oplus \varepsilon\}$$

where

$$A \oplus \varepsilon = \bigcup_{x \in A} \mathcal{B}(x, \varepsilon)$$

with $\mathcal{B}(x, \varepsilon)$ the ball of radius ε centered on x. Assume S to be r-dimensional and P to have its mass spread all over S. Formally: $\exists c > 0 \mid \forall x \in S, \varepsilon > 0, P(B(x, \varepsilon)) \geq c\varepsilon^r$ and the number of balls of size ε required to cover S is $C(\frac{1}{\varepsilon})^r$. Then:

$P(H(\widehat{S}, S) > 2\varepsilon)$
$= P$(there is no ball in $S \oplus \varepsilon$ with no sample point inside)
$\leq C\varepsilon^{-r} \max_k (1 - P(X_1 \in B_k))^n$
$\leq C\varepsilon^{-r}(1 - \min_k P(X_1 \in B_k))^n$
$\leq C\varepsilon^{-r} e^{-nc\varepsilon^r}$ (using $P(X \in B_k) \geq c\varepsilon^r$ and $(1-x)^n \leq exp(-nx)$ for $0 \leq x \leq 1$)
which converges.

Note that we only verify one of the two inclusions for the Hausdorff distance since the other one is always respected.

Now that we can find an estimation for a manifold that would concentrate all or most of a distribution. We try to estimate the topology of a manifold. That will be done in details with persistent homology however let us tackle the topic already. To estimate the topology of a set, one can try to find an estimator of the topology that is topologically similar to the studied set. Studying topological similarity comes down to the study of the existence of homeomorphism i.e. of bi-continuous maps. Markov [10] showed that in general the question is undecidable for dimension greater than four. That is why instead we determine if two sets are homologically equivalent.

A result in topology and statistics showed that:

$$\widehat{S} = \bigcup_{i=1}^{n} B(X_i, \varepsilon)$$

has the homology of S with high probabiity as long as S has a positive reach and ε remains relatively small compared to it. The reach r of a set S is the largest real number such that any point within distance r of S has a unique projection on S.

3.2 Persistent homology

Persistent homology is the computation side of homology theory. Persistent homology is the name given to an algorithm which aim is to compute the topological features of data. We have discussed in detail homology from a mathematical point of view in part **2** and have shown that for $n \geq 2$ an n-dimensional sphere has for

homology group: $H_n^{\Delta}(\mathcal{S}^n) \approx \mathbb{Z}$. To simplify vocabulary and notation we are now to be writting using betti numbers. Betti numbers of an n-dimensional topological space X counts the power of $\mathbb{Z}$ each group $(H_k^{\Delta}(X))_{0\leq k\leq n}$ is topologically equivalent to. In the case of the torus T, the betti number are $(1, 2, 1, 0)$. Recalling the formula to compute $H_k^{\Delta}(X) = \frac{\text{Ker}\delta_n}{\text{Im}\delta_{n+1}}$, since $C_{n+1} = \emptyset$, $\delta_{n+1} = 0$ and hence any value $(H_k^{\Delta}(X))_{k>n} = 0$. Therefore, it is conventionally admitted not to write betti numbers above that value.

4 Knot theory

4.1 Fundamental concepts

4.2 Knot determinant

4.2.1 Definition

4.2.2 Algorithms

5 Measuring neural network expressiveness

5.1 Using topological data analysis

5.2 Using trajectories

6 The study of trajectories from a knot theory perspective

6.1 Methodology

6.2 Algorithms

6.3 Results

7 Extending the study of expressiveness with topological data analysis

7.1 Methodology

7.2 Algorithms

7.3 Results

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, Dec. 1989.

[2] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, pp. 251–257, Jan. 1991.

[3] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding Deep Neural Networks with Rectified Linear Units," *arXiv:1611.01491 [cond-mat, stat]*, Feb. 2018. arXiv: 1611.01491.

[4] S. Wang and X. Sun, "Generalization of hinging hyperplanes," *IEEE Transactions on Information Theory*, vol. 51, pp. 4425–4431, Dec. 2005.

[5] B. Hanin and M. Sellke, "Approximating Continuous Functions by ReLU Nets of Minimal Width," *arXiv:1710.11278 [cs, math, stat]*, Mar. 2018. arXiv: 1710.11278.

[6] A. Hatcher, *Algebraic topology*. Cambridge ; New York: Cambridge University Press, 2002.

[7] P. Albin, "1. History of Algebraic Topology; Homotopy Equivalence - Pierre Albin," 2018.

[8] P. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, vol. 1, (Edinburgh, UK), pp. 958–963, IEEE Comput. Soc, 2003.

[9] L. Wasserman, "Topological Data Analysis," *arXiv:1609.08227 [stat]*, Sept. 2016. arXiv: 1609.08227.

[10] A. Markov, "Insolubility of the Problem of Homeomorphy," 2001.