Introduction to Neural Network Approximation Theory

Marcus Hutter

DeepMind, London, UK
http://www.hutter1.net/
Abstract

Artificial Neural Networks (NN) have achieved impressive performance on a wide range of tasks, especially in natural language processing and vision. Mathematically, NN represent function classes, leading to natural and important capacity questions: (a) which functions can a NN represent, (b) approximate arbitrarily well, (c) how large does a NN have to be, (d) does depth increase capacity. This tutorial will discuss (a)-(d) for the Multi-Layer Perceptron (MLP) which is the oldest and most successful NN architecture. In this endeavor I will also visit some classical mathematical representation and approximation theorems. Deep learning theory and effective=learning capacity are beyond the scope of this tutorial, but basic knowledge of (a)-(d) is important to appreciate these more sophisticated topics. The tutorial is mostly based on the classical paper by Allan Pinkus, but with illustrations, and proofs replaced by proof ideas.
# Table of Contents

1. Motivation/Preliminaries
2. Shallow Neural Networks
3. Universality=Density of 1HLP
4. Variations
5. Pathological Approximations
6. Degree of Continuous Approximation
7. Two Hidden Layer Perceptron (2HLP)
# Table of Contents

1. Motivation/Preliminaries
2. Shallow Neural Networks
3. Universality=Density of 1HLP
4. Variations
5. Pathological Approximations
6. Degree of Continuous Approximation
7. Two Hidden Layer Perceptron (2HLP)
Neural Network (NN)

One Hidden Layer

Two Hidden Layers
What does Universality of NN Mean?

- **Problem of density**: Can a sufficiently large NN approximate any reasonable function arbitrarily well? (which metric/norm/topology/domain, which function class)

- **Degree of approximation**: How well can a specific NN size approximate specific function classes (above + NN depth/width)

- **Interpolation**: Can (poly-size) NN exactly represent the finite data \( D = (x_1, y_1), \ldots, (x_T, y_T) \).

- **Representation/Approximation/Learning Capacity**: Size of function class that can be represented/approximated/learned.

- **Universal Function Approximator**: Something that can approximate any (continuous) function.
Why Care?

- NN are very popular and successful, but hard to understand, so every insight helps.
- Being able to approximate a function is a necessary pre-condition for being able to learn it.
- Some learning algorithms can sometimes find the global minimum. E.g. Stochastic Gradient Descent or Simulated Annealing. In this case Approximation = Learning capacity.
- Approximation capacity relevant for understanding overfitting and interpolation (phenomena).
- Is research on shallow NN exhausted? Little know about benefits of deep NN or non-MLP!
- Basis for capacity results of recent (anti)symmetric NN.
Why Mostly Pre-2000 Results

- Pinkus (1999) is a great 50-page review incl. proofs.
- My presentation essential follows Pinkus (1999) except:
  - Proof sketches/ideas instead of technical proofs.
  - Minor omissions/additions.
- Graphics/Images from Wikipedia, Internet, Myself [Wik].
- Why a 20 year-old paper?
- NN approximation theory research was most active pre-2000.
- You need to know some classics.
- It’s IMO still the single best paper on NN approximation theory.
- You can only de/appreciate newer work knowing Pinkus (1999).
Beyond the Scope of this Introduction

- Generalization
- Learning algorithms/capacity
- Deep NN
- Applications / Empirical studies
- Optimization theory
- Relation to SVM & Kernels & Gaussian Processes
- Other NN architectures (stochastic/spiking/adversarial)
- \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) function to be approximated by NN \( \Phi \).
- \( x \equiv (x_1, ..., x_n) \in \mathbb{R}^n \) input to NN, sometimes \( \in [0; 1]^n \)
- \( y \in \mathbb{R} \) output of NN

Pay attention to Definitions (red)
Table of Contents

1 Motivation/Preliminaries

2 Shallow Neural Networks

3 Universality=Density of 1HLP

4 Variations

5 Pathological Approximations

6 Degree of Continuous Approximation

7 Two Hidden Layer Perceptron (2HLP)
Simplest and oldest 1-layer NN model:

Thresholded linear function:

\[ y = \Phi(x) := \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{n} w_i x_i + b \geq 0 \\ 0 & \text{else} \end{cases} \]

\( w_i \in \mathbb{R} \) are synaptic weights, \( b \in \mathbb{R} \) is bias.
Perceptron (1958)

- 20×20 pixel camera input
  = 400 photocells
- Weights = potentiometers
- Weight updates by electric motors

The New York Times:
”[The perceptron] is the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.”
Can represent all functions that are 1 in some half-space of $\mathbb{R}^n$ and 0 in the complement half-space.

Can be used to classify linearly separable data

$$D := \{(x_1, y_1), \ldots, (x_T, y_T)\} \equiv \{(x_t, y_t) : 1 \leq t \leq T\}$$

**Learnable: Perceptron: Iterate**

$$w \leftarrow w - \eta(y_t - f(x_t))x_t$$

**But:** This talk is not concerned about learnability, but only Representation

Representation is necessary but not sufficient for learnability
McCulloch-Pitts - Limitations

- Can represent only binary functions $y \in \{0, 1\}$.
- Discontinuous and non-differentiable, indeed $\Phi$ is piecewise constant. Hence it cannot (directly) be learnt by gradient descent.
- Not universal, e.g. cannot represent XOR function. Pointed out by Marvin Minsky: Caused first NN winter.
- But: Perceptron + KernelTrick = conceptual foundations of Support Vector Machines (SVMs).
One Neuron Perceptron

- \( y = \Phi(x) := \sigma(\sum_{i=1}^{n} w_i x_i + b) \equiv \sigma(w \cdot x + b) \),

- \( \sigma : \mathbb{R} \rightarrow \mathbb{R} \) activation function (examples next slide).

- Generalizes McCulloch-Pitts: \( \sigma(x) = 1 \) if \( x \geq 0 \) else 0.

- \( \Phi \) is continuous/smooth if \( \sigma \) is continuous/smooth.

- **Universal (useless) interpolator:** \( \forall D \ \exists \tilde{\sigma}, w, b : \Phi(x_t) = y_t \ \forall t \leq T \)
  Proof: Choose \( w \) randomly, then all args. of \( \tilde{\sigma} \) differ (true for most \( w \))
  Even \( \exists \tilde{\sigma} \forall D \forall \varepsilon > 0 \exists w, b : |\Phi(x_t) - y_t| < \varepsilon \ \forall 1 \leq t \leq T \)

- **Problem:** \( \tilde{\sigma} \) is pathological (more later)

- **Limitation:** Can only model fcts constant in all-but-one direction (\( w \))
  e.g. cannot even model \( f(x) = x^2 + y^2 \) (but \( \sigma = \sin^2 \) can model XOR!)
Historical/Popular Activation Functions

- **STEP:** \( \sigma(x) = 1 \) if \( x \geq 0 \) else 0 (Heaviside, McCulloch-Pitts)
- **SIGMOID:** \( \sigma(x) = 1/(1 + e^{-x}) \) logistic sigmoid (bounded, smooth)
- **TANH:** \( \sigma(x) = \tanh(x) \) ”signed” sigmoid (bounded, smooth)
- **ReLU:** \( \sigma(x) = \max\{x, 0\} \) rectified linear unit (simple, good \( \nabla \) for \( x > 0 \))
- **ARCTAN:** \( \sigma(x) = \arctan(x) \) \( (\sigma'(x) \rightarrow 0 \) slowly for \( x \rightarrow \infty \))
- **HARD-TANH:** \( \sigma(x) = \min\{1, \max\{x, -1\}\} \) (bounded, simple)
- **LEAKY-ReLU:** \( \sigma(x) = \max\{x, 0.01x\} \) (avoids \( \sigma' = 0 \))
- **SMOOTH-ReLU:** \( \sigma(x) = \log(1 + \exp(x)) \) (smooth, good \( \nabla \) for \( x > 0 \))
- **LOGIT:** \( \sigma(x) = \log(x/(1 - x)) \) (map prob: \( (0; 1) \rightarrow \mathbb{R} \), inv.SIGMOID)
- **POLY:** \( \sigma(x) = x^2 \) or higher polynomial (bad for shallow NNs)
- **SOFTMAX:** \( \sigma(x_1, \ldots, x_n) = e^{x_i}/\sum_{i=1}^{n} e^{x_i} \) (output probability vector)

**SIGMOID** is all-time favorite. **ReLU** is current favorite.
Historical/Popular Activation Functions

\[ \sigma(x) \]

- **STEP**
- **SIGMOID(4x)**
- **TANH**
- **ReLU**
- **ARCTAN**
- **HARD-TANH**
- **LEAKY-ReLU**
- **SMOOTH-ReLU**
- **LOGIT/4**
- **SQUARE/2**

![Activation Functions Graph](image-url)
Desirable Properties of Activation Functions $\sigma$

- **Simple** (for speed)
- **Monotone** (avoid misleading gradients)
- **Bounded** (to keep activation ranges small in Deep NN)
- **(Sub)Differentiable** (for Gradient Descent)
- **Smooth** (to represent smooth functions, e.g. required in physics)
- **Gradient does not vanish** too quickly for large input

Leads to *universal* approximator in NNs: we will see, this is a very mild condition, even for shallow NN
One-Hidden-Layer Perceptron (1HLP)

\[ y = \Phi(x) := \sum_{j=1}^{r} c_j \sigma(\sum_{i=1}^{n} w_{ji} x_i + b_j) \equiv c \cdot \sigma(Wx + b) \]

- The hidden layer \( \sigma(W \cdot + b) \) is non-linear
- The output layer \( c \cdot \) is linear
- One could apply another activation function to the output layer
- This usually does not increase capacity, sometimes it even decreases it
- The 1HLP model is already a universal function approximator for nearly any choice of \( \sigma \) (we will show)
- Obvious extension to Multi-Layer Perceptron (MLP): Discussed later.
Table of Contents

1 Motivation/Preliminaries
2 Shallow Neural Networks
3 Universality = Density of 1HLP
4 Variations
5 Pathological Approximations
6 Degree of Continuous Approximation
7 Two Hidden Layer Perceptron (2HLP)
Using 1HLP for Classification

- Heaviside activation function: \( \sigma(x) = 1 \) if \( x \geq 1 \) else 0
- McCulloch-Pitts model \( y = \sigma(w \cdot x + b) \) could not represent XOR.
- Can \( T \) points \( x_t \in \mathbb{R}^n \) be separated=classified by 1HLP?
- Early result by Baum (1988): \( r = T/n \) neurons suffice.
- And are needed for some, e.g. for XOR.

**Theorem**

A 1HLP can perfectly classify any ‘general’ \( D \in (\mathbb{R}^n \times \{0, 1\})^T \) if and only if the 1HLP has \( r = \lceil T/n \rceil \) (or more) hidden neurons.

The mild ‘general’ conditions are:

- \((x_t, 1) = (x_s, 0)\) only if \( x_t \neq x_s \) (obviously necessary), and
- no \( n \) data points are linearly dependent (randomize infinitesimally)
For $|\mathcal{Y}| > 2$ class labels, reduce the problem to $\lceil \log |\mathcal{Y}| \rceil$ binary classification problems: $r = \lceil \log |\mathcal{Y}| \rceil \cdot \lceil T/n \rceil$.

Examples: $r$ neurons suffice to perfectly classify:

| Data Set      | $T$   | $n$    | $|\mathcal{Y}|$ | $r$  |
|---------------|-------|--------|----------------|-----|
| MNIST         | 70'000| 28x28  | 10             | 360 |
| CIFAR10       | 60'000| 32x32x3| 10             | 80  |
| CIFAR100      | 60'000| 32x32x3| 100            | 140 |
| ImageNet      | $14\times10^6$ | 256x256x3 | 21'000        | 1’080 |

Result also true for most other $\sigma$:

\[
\text{SIGMOID}(x/\varepsilon) \approx \text{STEP}(x) \approx [\text{ReLU}(x + \varepsilon) - \text{ReLU}(x)]/\varepsilon
\]

Result very recently extended to regression [?]
Constructive Proof (Sketch) of 1HLP Upper Bound for Classification

- Let \( D^+ := \{(x, y) \in D : y = 1\} \).
- W.l.o.g. assume \(|D^+| \leq T / 2\).
- Partition \( D^+ \) in groups of \( n \) points.
- For each group, choose hyperplane \( \mathbf{w} \cdot \mathbf{x} + b \) through \( n \) points.
- Choose pair of neurons:
  \[ \text{STEP}(\mathbf{w} \cdot \mathbf{x} + b + \varepsilon) - \text{STEP}(\mathbf{w} \cdot \mathbf{x} + b - \varepsilon). \]
- On \( D \) this is only 1 for the \( n \) points.
- Add up all \( \leq \lceil T / 2n \rceil \) such pairs of neurons in output layer.
Which Functions can 1HLP Represent?

\( \mathcal{M}_r(\sigma) := \{c \cdot \sigma(Wx + b) : b, c \in \mathbb{R}^r, W \in \mathbb{R}^{r \times n}\} \)

The set of all functions exactly \textit{representable} by a one-hidden-layer perceptron (1HLP) with \( r \) hidden neurons.

\( \mathcal{M}(\sigma) := \text{span} \{\sigma(w \cdot x + b) : w \in \mathbb{R}^n, b \in \mathbb{R}\} \equiv \bigcup_{r=1}^{\infty} \mathcal{M}_r(\sigma) \)

Set of all fcts exactly \textit{representable} by a 1HLP of arbitrary \textit{width} \( r \)

Let \( \mathcal{C}(\mathbb{R}^n) \) be the set of \textit{continuous functions} from \( \mathbb{R}^n \) to \( \mathbb{R} \)

If not mentioned otherwise we will in the following assume that \( \sigma \) \textit{is continuous}, i.e. \( \sigma \in \mathcal{C}(\mathbb{R}) \).

For such \( \sigma \), all 1HLP are continuous functions, i.e. \( \mathcal{M}(\sigma) \subseteq \mathcal{C}(\mathbb{R}^n) \).

But 1HLP cannot \textit{represent} all continuous functions, i.e. \( \mathcal{M}(\sigma) \neq \mathcal{C}(\mathbb{R}^n) \).

Proof: If \( \sigma \) is differentiable, then all \( \Phi \in \mathcal{M}(\sigma) \) are differentiable.
Which Functions can $1$HLP Approximate?

- Can $M(\sigma)$ approximate every continuous function?
- Functions can be approximated w.r.t. different topologies/metrics.

**Definition (Convergence Uniformly on Compacta (CUC))**

$f_n \in C(\mathbb{R}^n)$ is said to Converge Uniformly on Compacta to $f \in C(\mathbb{R})$ ($f_m \xrightarrow{\text{CUC}} f$) iff $\forall \varepsilon > 0 \ \forall$ compact $K \subset \mathbb{R}^n \ \exists m_{\varepsilon,K} \in \mathbb{N} \ \forall m > m_{\varepsilon,K} : \max_{x \in K} |f_m(x) - f(x)| < \varepsilon$

- CUC corresponds to the compact-open topology e.g. induced by norm $\|f\|_{\text{CUC}} := \sup_{k \in \mathbb{N}} k^{-2} \sup_{x \in [-k;k]^n} |f(x)|/(1 + \sup_{x \in [-k;k]^n} |f(x)|)$.
- This is a very strong notion of convergence. CUC implies convergence in $L^p(K, \mu)$ for any $1 \leq p \leq \infty$, and compact $K$, and any nonnegative finite Borel measure $\mu$ on $K$. 
Universality of 1HLP

- Let $\overline{\mathcal{M}(\sigma)}$ be the closure of $\mathcal{M}(\sigma)$ w.r.t. compact-open topology, i.e. $\overline{\mathcal{M}(\sigma)}$ is the set of all functions that can be approximated arbitrarily well by a sufficiently wide 1HLP.

- Let $\mathcal{M}_\infty(\sigma)$ be the set of functions representable by an infinite 1HLP.

- Exercise: Is $\overline{\mathcal{M}(\sigma)} = \mathcal{M}_\infty(\sigma)$?

- A key result in NN approximation theory is that 1HLP can approximate every continuous function for most $\sigma$:

Theorem (Universality of one-hidden-layer perceptron)

Let $\sigma \in \mathcal{C}(\mathbb{R})$. Then $\overline{\mathcal{M}(\sigma)} = \mathcal{C}(\mathbb{R}^n)$ iff $\sigma$ is not a polynomial.

- Many proofs of (variations of) this result: First one by L. Schwartz (1944)!
- Only-if is easy: If $\sigma$ is poly of degree $d$, then $\mathcal{M}(\sigma)$ only contains all multivariate polys of at most degree $d$, which are not dense in $\mathcal{C}(\mathbb{R}^n)$. 
Density/Approximation/Universality
Proof Techniques

- Discretized inverse *Radon transform*

- *Hahn Banach theorem* and *Riesz Representation theorem* (continuous linear functionals on the space of continuous functions)

- *Stone-Weierstrass Theorem* (we will use)

- *Ridge functions*: Reduces the problem to the univariate case

- *Kolmogorov-Arnold representation theorem*:
  Exact representation for finite 2HLP, but *pathological* ς.

- Other pathological *tabulation* and *binarization* methods, e.g. [LSYZ20]
Weierstrass Approximation Theorem

Every continuous function can be approximated by a polynomial:

**Theorem (Weierstrass Approximation)**

\[ \forall f \in C([a; b]) \, \forall \varepsilon > 0 \, \exists \text{ polynomial } p \, \forall x \in [a; b]: \, |f(x) - p(x)| < \varepsilon \]

Proof: Convolve \( f \) with polynomial mollifier \( p_n \) makes it poly. \( p = f \ast p_n \)

\[ f(x) \ast p_n(x) = (1-x^2)^n \quad [n=100] \]

\[ (p_n \ast f)(x) \quad [n=100] \]
Proof-Sketch of Weierstrass Theorem

- Scale domain to $[0; 1]$ and tilt $f$ to be 0 at boundary:
  Define $g(t) := f(a + t(b - a)) - f(a) - t(f(b) - f(a))$ for $t \in [0; 1]$ and 0 outside $[0; 1]$.

- $g$ is continuous and $g(0) = g(1) = 0$.

- If we can approximate $g$ by a polynomial, then clearly also $f$.

- A mollifier $p_n(x)$ is a smooth function sharply peaked at 0 such that $\int p_n(x) dx = 1.$ and $(p_n * g)(x) := \int p_n(t) g(x - t) dt \approx g(x)$. Assume $p_n$ tends to the Dirac $\delta$ for $n \to \infty$.

- If $p_n$ is a polynomial, then $p_n * g$ is also a polynomial.

- Polynomial $p_n(x) = c_n(1 - x^2)^n$ on $[-1; 1]$ has this property.

- Crucial: $p_n(x)$ for $x \not\in [-1; 1]$ not “used”, since $g = 0$ outside $[0; 1]$.

- One can show $p_n * g \to g$ uniformly.

\[\square\]
Definition (separating points)

A set $A$ of functions defined on $X$ is said to separate points if for every two different points $x$ and $y$ in $X$ there exists a function $p$ in $A$ with $p(x) \neq p(y)$.

- Obviously if for some points $x \neq y$, all functions $p \in A$ have $p(x) = p(y)$, then no algebraic combination of such functions can have different values on $x$ and $y$.
- So separation is a necessary condition for representing all continuous functions. It turns out that this necessary condition is also sufficient:

Theorem (Stone-Weierstrass)

Suppose $X$ is a compact Hausdorff space (e.g. $[0, 1]^d$) and $A$ is a sub-algebra of $C(X)$ which contains a non-zero constant function. Then $A$ is dense in $C(X)$ if and only if it separates points.
$\sqrt{t}$ can be arbitrarily well approximated by polynomials on $[0, 1]$. Direct proof: The iteration $w(t) \leftarrow w(t) + \frac{1}{2}(t - w^2(t))$ (starting from $w(t) = 0$) converges to $\sqrt{t}$ and all iterates are polynomials.

This implies $|t| = \sqrt{t^2}$ and hence $2 \max\{t, s\} = |t - s| + t + s$ are approximable.

Hence $\min\{t_1, ..., t_n\}$ and $\max\{t_1, ..., t_n\}$ are approximable.

Assume we want to approximate $f : X \to \mathbb{R}$.

Assume $h(x)$ separates $a \in X$ and $b \in X$.

Use it to construct $g_{ab}(x)$ such that $g_{ab}(a) = f(a)$ and $g_{ab}(b) = f(b)$.
(Roughly) take sufficiently fine finite subset $X' \subseteq X$.

Then $g_a(x) := \min_{b \in X'} g_{ab}(x) \lesssim f(x)$ and $g_a(a) = f(a)$.

Then $g(x) := \max_{a \in X'} g_a(x) \gtrsim f(x)$ since $x' \in X' : g(x') \geq f(x')$.

Since also $g(x) \lesssim f(x)$, we get $g(x) \approx f(x)$.  

How is Stone-Weierstrass used in Proving Density of NN?

1. Allow sums and products of activation functions.

2. This permits to apply Stone-Weierstrass to obtain density.

3. Prove desired result without products, using (co)sine functions and the ability to write products of (co)sines as linear combinations of (co)sines [HSW89].

4. Or directly show that smooth $\sigma$ can approximate monomials, hence polynomials (later)
Ridge Functions

- Functions $g : \mathbb{R}^n \rightarrow \mathbb{R}$ of the form $g(a_1 x_1 + \ldots + a_n x_n) \equiv g(a \cdot x)$

- $a = (a_1, \ldots, a_n) \in \mathbb{R}^n \setminus \{0\}$ is a fixed direction.

- $g$ is constant on parallel hyperplanes orthogonal to $a$.

Density of Ridge Functions

- \( \mathcal{R}[G] := \text{span}\{g(a \cdot x) : a \in \mathbb{R}^n, g \in G \subseteq \mathbb{R} \rightarrow \mathbb{R}\} \).

- Obviously \( \mathcal{R}[G] \supseteq \mathcal{M}(\sigma) \) if \( G \supseteq \{\sigma(t + b) : b \in \mathbb{R}\} \) \( (t \in \mathbb{R}^1) \)

**Theorem (Ridge functions can approximate all continuous functions)**

\[ \overline{\mathcal{R}[C(\mathbb{R})]} = C(\mathbb{R}^n) \], i.e. \( \mathcal{R}[C(\mathbb{R})] \) is (CUC-)dense in \( C(\mathbb{R}^n) \).

One can already show that \( \mathcal{R}[G] \) is dense in \( C(\mathbb{R}^n) \) for much smaller \( G \):

- \( G = \{\sin, \cos\} \) (by Fourier transform),

- \( G = \{\exp\} \) (by bilateral Laplace transform),

- \( G = \{t^k, k \in \mathbb{N}_0\} \) (by some multiv. polynomial repr. theorem).

- \( G \supseteq \{\sigma(x + b) : b \in \mathbb{R}\} \) if \( \sigma \) is not a poly. (by earlier density thm.)
Reduction to One-Dimensional Case

- $N_r(\sigma) := \{ \sum_{i=1}^{r} c_i \sigma(\lambda_i t + \vartheta_i) : c_i, \lambda_i, \vartheta_i \in \mathbb{R} \} \equiv M_{r=1}^{n=1}(\sigma) \quad (t \in \mathbb{R}^1)$
- $N(\sigma) := \text{span}\{ \sigma(\lambda t + \vartheta) : \lambda, \vartheta \in \mathbb{R} \} \equiv \bigcup_{r=1}^{\infty} N_r(\sigma) \equiv M_{n=1}^{n=1}(\sigma)$
- $\mathcal{R}[N_1(\sigma)] = \mathcal{R}[N_r(\sigma)] = \mathcal{R}[N(\sigma)] = M(\sigma)$

Theorem (Reduction of density to one-dimensional case)

If $\overline{N(\sigma)} = \mathcal{C}(\mathbb{R})$ then $\overline{M(\sigma)} = \mathcal{C}(\mathbb{R}^n)$

$\implies$ Can focus on one-dimensional case! Great simplification.

Proof idea:

- Use Ridge Theorem to approximate $f : \mathbb{R}^n \to \mathbb{R}$ as mixture of $r$ continuous $g_i : \mathbb{R} \to \mathbb{R}$, i.e. $f \approx \in \mathcal{R}\{g_1, \ldots, g_r\}$.

- Now $g_i \approx \in N_{m_i}(\sigma)$ by assumption on $N(\sigma)$.

- Combining both to one linear approx. shows $f \approx \in M_{m_1+\ldots+m_r}(\sigma)$. 
Let $C^\infty(\mathbb{R})$ be the class of all $\infty$-often differentiable functions $f : \mathbb{R} \to \mathbb{R}$.

**Theorem (Universality of 1d 1HLP for most smooth $\sigma$)\)**

If $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial, then $\overline{\mathcal{N}(\sigma)} = \mathcal{C}(\mathbb{R})$. Furthermore $\overline{\mathcal{N}_r(\sigma)}$ includes all polynomials of degree $< r$.

**Proof:**

- **Exercise:** Since $\sigma$ is not a polynomial, there exists $\vartheta_0$ for which all derivatives $\sigma^{(k)}(\vartheta_0) \neq 0$.

- $\sigma((\lambda + \varepsilon)t + \vartheta_0) - \sigma((\lambda - \varepsilon)t + \vartheta_0) \in \mathcal{N}_2(\sigma)$, hence $t\sigma'(\vartheta_0) \equiv d\sigma(\lambda t + \vartheta_0)/d\lambda|_{\lambda=0} \in \mathcal{N}_2(\sigma)$.

- Induction shows $t^k\sigma^{(k)}(\vartheta_0) \equiv d^k\sigma(\lambda t + \vartheta_0)/d\lambda^k|_{\lambda=0} \in \mathcal{N}_{k+1}(\sigma)$.

- Hence all monomials, hence all polynomials $\in \overline{\mathcal{N}(\sigma)}$.

- Hence by Weierstrass Theorem $\overline{\mathcal{N}(\sigma)} = \mathcal{C}(\mathbb{R})$. \qed
Weaker Assumptions on $\sigma$

Assumption $\sigma \notin \text{Poly}$ was necessary and cannot be dropped)

- $\sigma \in C^\infty([a; b])$ for some interval $(a < b)$ (same proof)
- $\sigma \in C(\mathbb{R})$. Proof idea: Mollify $\sigma \approx \sigma \phi := \sigma \ast \phi \in C^\infty(\mathbb{R}) \cap \overline{N(\sigma)}$.
- $\sigma$ bounded and Riemann-integrable on every finite interval. Proof idea: Same mollifier idea + approx. $\int$ in $\ast$ by $\sum$ to show $\sigma \phi \in \overline{N(\sigma)}$
- $\sigma$ bounded and Riemann-integrable on $[a; b]$ (combine proofs)
- $\sigma \in C(\mathbb{R}) \cap L^1(\mathbb{R})$ then $\overline{N(\sigma)}|_{\lambda=1} = C(\mathbb{R})$ (proof based on Fourier transform)

Remark: Results remain valid if input $x$ is preprocessed by continuous injection.
Some applications require not only to approximate the function well, but also its derivatives (e.g. in physics).

**Multivariate derivatives**: For \( m \equiv (m_1, \ldots, m_n) \in \mathbb{N}_0^n \) and 
\[
|m| := m_1 + \ldots + m_n \quad \text{and} \quad x^m := x_1^{m_1} \cdots x_n^{m_n}
\]
let 
\[
D^m := \frac{\partial |m|}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}.
\]

Differentiable functions: 
\[
C^m(\mathbb{R}^n) := \left\{ f : D^k f \in C(\mathbb{R}^n) \ \forall k \leq m \right\}
\]
where \( k \leq m :\leftrightarrow k_i \leq m_i \forall i \). 
\[
C^{m_1, \ldots, m_s}(\mathbb{R}^n) := \bigcap_{i=1}^s C^{m_i}(\mathbb{R}^n).
\]
\[
C^m(\mathbb{R}^n) := \bigcap_{|m|=m} C^m(\mathbb{R}^n) = \left\{ f : D^k f \in C(\mathbb{R}^n) \ \forall |k| \leq m \right\}.
\]

**CUC^m**: We say \( \mathcal{M}(\sigma) := \text{span}\{\sigma(w \cdot x + b) : w \in \mathbb{R}^n, b \in \mathbb{R}\} \) is dense in \( C^{m_1, \ldots, m_s}(\mathbb{R}^n) \) if, for any \( f \in C^{m_1, \ldots, m_s}(\mathbb{R}^n) \), any compact \( K \subset \mathbb{R}^n \), any \( \varepsilon > 0 \), there exists \( g \in \mathcal{M}(\sigma) \) satisfying 
\[
\max_{x \in K} |D^k f(x) - D^k g(x)| < \varepsilon \quad \text{for all} \ k \in \mathbb{N}_0^n \ \text{for which} \ \exists i : k \leq m^i
\]
Blown-up definitions and proofs. *Little new insight*
Theorem (1HLP is dense in $C^m$)

Let $m^i \in \mathbb{N}_0^n$ and $m := \max\{|m^i| : i = 1, \ldots, s\}$. Assume $\sigma \in C^m(\mathbb{R})$ and $\sigma$ not polynomial. Then $\mathcal{M}(\sigma)$ is $(\text{CUC}^m)$-dense in $C^{m_1, \ldots, m_s}(\mathbb{R}^n)$.

Proof idea:
- Exercise: Multivariate polynomials are dense in $C^{m_1, \ldots, m_s}(\mathbb{R}^n)$, so it suffices to approximate polynomials.
- Exercise: Any multivariate polynomial $h$ can be represented as $h(x) = \sum_{i=1}^{r} p_i(a^i \cdot x)$, where $p_i$ are univariate polynomials (mentioned and used before).
- Therefore we only need to approximate univariate polynomials.
- Approximate the $m$-th derivative of $p_i$ and then integrate. If $p_i^{(m)} \approx f_i^{(m)} \in \mathcal{N}(\sigma^{(m)})$, then also for integrals on compacta $p_i^{(k)} \approx f_i^{(k)} \in \mathcal{N}(\sigma^{(k)}) \forall k < m$. 

Marcus Hutter
Universality of Neural Networks
DeepMind 42 / 77
Interpolation vs Approximation

Interpolation

Crude Approx.

Approximation

Data
Interpolation by 1HLP

**Theorem (1HLP with $T$ Neurons can Interpolate $T$ data items)**

For any $\sigma \in C(\mathbb{R}) \setminus Poly$ and any $D = (x_1, y_1), \ldots, (x_T, y_T)$, there exists NN $\Phi \in \mathcal{M}_T(\sigma)$ (1HLP with $T$ neurons) such that $\Phi(x_t) = y_t$ for all $1 \leq t \leq T$.

- Interpolation is *different* from approximation.
- **Harder**: Asks for exact representation at finitely many points.
- **Easier**: No constraint on NN outside of data points.
- In ML we want to generalize to new data rather than interpolate.
- But minimizing empirical loss leads to interpolation.
- Sometimes even interpolating NN can generalize well [Bel18].
- Hence: Interpolation questions/results are also (somewhat) interesting.
Proof Idea

- Reduce to one-dimensional 1HLP:
  Choose projection direction $\mathbf{v}$ so that all $z_t := \mathbf{v} \cdot \mathbf{x}_t$ are all different. (Always possible. Proof: random direction works w.p.1)

- Choose $\mathbf{w}_i = \lambda_i \mathbf{v}$, then $\mathcal{M}_T(\sigma)$ reduces to $\mathcal{N}_T(\sigma)$

- Need to show $\exists \phi \in \mathcal{N}_T(\sigma) : \phi(z_t) \equiv \sum_{j=1}^{T} c_j \sigma(\lambda_j z_t + \vartheta_j) = y_i$ for $1 \leq t \leq T$.

- Suffices to prove that $\sigma(\lambda z_t + \vartheta)$ are linearly independent functions of $\lambda$ and $\vartheta$ for $t = 1, \ldots, T$.

- $\sigma(\lambda \cdot + \vartheta)$ span $\mathcal{C}(\mathbb{R})$, hence (by some fancy argument) $\sigma(\cdot z_t + \cdot)$ are independent.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Motivation/Preliminaries</td>
</tr>
<tr>
<td>2</td>
<td>Shallow Neural Networks</td>
</tr>
<tr>
<td>3</td>
<td>Universality=Density of 1HLP</td>
</tr>
<tr>
<td>4</td>
<td>Variations</td>
</tr>
<tr>
<td>5</td>
<td>Pathological Approximations</td>
</tr>
<tr>
<td>6</td>
<td>Degree of Continuous Approximation</td>
</tr>
<tr>
<td>7</td>
<td>Two Hidden Layer Perceptron (2HLP)</td>
</tr>
</tbody>
</table>
Theorem (Universal pathological approximation by stitching)

There is a single (pathological) $\tilde{\sigma} \in C^\infty(\mathbb{R})$ that can approximate every continuous $f : [0; 1] \to \mathbb{R}$ by translation:

$$\forall \varepsilon > 0 \ \forall f \in C[0; 1] \ \exists m \in \mathbb{N} : |\tilde{\sigma}(x + m) - f(x)| < \varepsilon \ \forall x \in [0; 1].$$

Proof idea: Stitch together all polynomials with rational coefficients:
Pathological Proof

- Every $f \in C[0; 1]$ can be approximated by a polynomial with rational coefficients.

- Let $p_0, p_1, p_2, \ldots \in C[0; 1]$ be some enumeration of the countably many such polynomials.

- $\forall m \in \mathbb{N}_0$ define $\tilde{\sigma}(z + 2m) := p_m(x)$ for $z \in [0; 1]$ and interpolate $\tilde{\sigma}$ smoothly between $2m + 1$ and $2m + 2$.

- By construction $\tilde{\sigma}$ is smooth.

- Let $m$ be such that $|p_m(z) - f(z)| < \varepsilon$. Then $|\tilde{\sigma}(z + 2m) - f(z)| < \varepsilon$.

In what follows we denote such pathological $\sigma$ by $\tilde{\sigma}$. 
Construction can be extended to $f \in C(\mathbb{R})$ and CUC-norm: Represent poly $p_m \in C[-k; k]$ for all $m, k \in \mathbb{N}$ in $\sigma(z + d(m, k)) := p_m(z)$ via suitable dovetailing $d$.

One can even choose $\tilde{\sigma}$ monotone increasing by tilting $\tilde{\sigma}$ (details later).

Some results in NN approximation theory use such pathological approximation.

Most are based on sophistications of stitching, but some are even worse.

For instance [LSYZ20] constructs NN essentially predicting the $k$-th bit of binary expansion of $f$, and stitch everything together maintaining even continuity.
Sobolev Space & Norm

- **Unit closed ball** in \( \mathbb{R}^n \): \( B^n := \{ x : \| x \|_2 \equiv (x_1^2 + ... + x_n^2)^{1/2} \leq 1 \} \)

- \( C^m(B^n) := \{ f : B^n \rightarrow \mathbb{R} : D^k f \subseteq \mathbb{R}^n \rightarrow \mathbb{R} \text{ defined & continuous} \quad \forall k : |k| \leq m \} \)

- **p-norm**: \( \| g \|_p := \begin{cases} (\int_{B^n} |g(x)|^p d\mathbf{x})^{1/p}, & 1 \leq p < \infty \\ \text{ess sup}_{x \in B^n} |g(x)|, & p = \infty \end{cases} \)

- **Sobolev norm**: \( \| f \|_{m,p} := \begin{cases} \sum_{0 \leq |k| \leq m} \| D^k f \|_p^{1/p}, & 1 \leq p < \infty \\ \max_{0 \leq |k| \leq m} \| D^k f \|_\infty, & p = \infty \end{cases} \)

- **Sobolev space**: \( \mathcal{W}_p^m \equiv \mathcal{W}(B^n) = \text{completion of } C^m(B^n) \text{ w.r.t. Sobolev norm} \)

- \( B_p^m \equiv B_p^m(B^n) := \{ f : f \in \mathcal{W}_p^m, \| f \|_{m,p} \leq 1 \} \quad = \text{set of functions on } B^n \text{ of bounded Sobolev norm} \)
Approximation in \( p \)-Norm

- \( B^n \) is compact, hence \( \mathcal{C}(B^n) \) is dense in \( L^p \equiv L^p(B^n) := W^0_p(B^n) \)

- For \( \sigma \in \mathcal{C}(\mathbb{R}) \setminus \text{Poly} \), \( \mathcal{M}(\sigma) \) is dense in \( \mathcal{C}(B^n) \) hence dense in \( L^p \)
Pathological Approximation Rates of 1HLP

**Theorem (Lower bound on approximation rate of 1HLP)**

For $n \geq 2$ and $m \geq 1$ and each $r \in \mathbb{N}$ and any $\sigma$, there exists $f \in B_2^m$ for which

$$\inf_{\Phi \in \mathcal{M}_r(\sigma)} \| f - \Phi \|_{L^2(B^n)} \geq C_{n,m} r^{-m/(n-1)}$$

- **Curse of dimensionality:** Error $\varepsilon \geq (1/r)^{1/(n-1)} \Rightarrow r \geq (1/\varepsilon)^{n-1}$
- The lower bound is attained for “most” functions $f$ (Maiorov 1999)
- Proof: difficult and complicated. See Maiorov (1999)

**Theorem (Upper bound on approximation rate of 1HLP)**

There exist sigmoidal and strictly increasing $\tilde{\sigma} \in C^\infty(\mathbb{R})$ for which for $n \geq 2$ and $m \geq 1$ and each $r \in \mathbb{N}$ and all $p \in [1; \infty]$ and all $f \in B_p^m$, we have

$$\inf_{\Phi \in \mathcal{M}_r(\tilde{\sigma})} \| f - \Phi \|_{L^p(B^n)} \leq C_{n,m} r^{-m/(n-1)}.$$  

- **Blessing of smoothness:** $\varepsilon \leq (1/r)^m \Rightarrow r \leq (1/\varepsilon)^{1/m}$
Theorem (Approximation Rate of Multivariate Polynomials)

Multivariate polynomials $P_k$ of degree at most $k$ can approximate any $f \in B_{p}^{m}$ to accuracy $O(k^{-m})$ in $p$-norm.

There even exists a linear operator $L : \mathcal{W}_p^m \rightarrow \mathcal{P}_k$ that finds the approximating polynomial, i.e. $\|f - L(f)\|_p \leq Ck^{-m}$.

Proof: Mhaskar (1996)
Proof Sketch of Pathological Upper Bound

- The vector space of $n$-variate polynomials $\mathcal{H}_k$ of exactly degree $k$ has dimension $r := \binom{n-1+k}{k} \approx k^{n-1}$ for $k \gg n$.

- A linear combination of $r$ ridge functions based on 1d polynomials of degree at most $k$ can represent all multivariate polynomials $P_k$.

- Any ridge functions can be approximated by one neuron to any accuracy $\varepsilon$.
  Proof: Construct and use pathological $\tilde{\sigma}$ similar as above in the 1d case, then lift via ridge functions to $n$-dim $\tilde{\sigma}(a \cdot x + b)$.

- By linear trafo one can even make each polynomial monotone increasing and stitch them overall together in a monotonically increasing way, and correcting the output with $n+1$ compensating linear transformation by defining some linear regions in $\tilde{\sigma}$ itself.

- Together this shows that $f \in B_2^m$ can be approximated by $\Phi \in M_{r'}(\tilde{\sigma})$ to accuracy $k^{-m} \approx r^m/(n-1) \approx r'^m/(n-1)$. 

Quiz: Do there exist continuous bijections $\beta : X \rightarrow Y$ that are not homeomorphisms?

Answer: If $X$ is compact and $Y$ is Hausdorff then not.

If $X$ is not compact, then it can happen. E.g. $\beta : [0; 2\pi) \rightarrow \mathbb{S}$.

This is the key “loophole” exploited by / problem with pathological stitching $\sigma$.

But $\sigma$ is a more interesting pathology (next slide)
Theorem (Dense Pathological Injections)

There are continuous bijections \( \tilde{\varphi} : [0; \infty) \to \text{Image}(\tilde{\varphi}) \) with \( \text{Image}(\tilde{\varphi}) \) dense in \( C[0; 1] \), but inverse \( \tilde{\varphi}^{-1} \) cannot be continuous.

Proof sketch of injectivity:

- Choose a unique enumeration \( \mathbb{N} \to \mathbb{Q}^* \cong \) rational polynomials.

- Choose \( \tilde{\sigma} \) as before but connect polynomials with distinct non-polynomials.

- Define \( \tilde{\varphi}(z) := \sigma_z \) with \( \sigma_z : [0; 1] \to \mathbb{R} \) with \( \sigma_z(x) := \tilde{\sigma}(x + 2z) \).

- \( \tilde{\varphi}(n + x) \neq \tilde{\varphi}(m + x) \) for \( \mathbb{Z} \ni n \neq m \in \mathbb{Z} \), since polys are different.

- \( \tilde{\varphi}(n + x) \neq \tilde{\varphi}(m + y) \) for \( x - y \notin \mathbb{Z} \), since break location differs.
Proof of non-continuity of $\check{\phi}^{-1}$:

- Consider polynomial $\sigma_0$ with some rational coeff. $a_0 \in \mathbb{Q}^m$.

- There is a sequence of rational vectors $a_k \neq a_0$ but $a_k \to a_0$.

- Let $n_k$ be the index of polynomial with coefficients $a_k$ (note $n_0 = 0$).

- Example: $m = 1$, $\sigma_{n_k} = a_k$, $a_0 = 0$ and $a_k = 1/k$.

- Then $\check{\phi}(n_k) \to \check{\phi}(0)$ but $n_k \to \infty \neq 0 = n_0$, hence $\check{\phi}^{-1}$ is not continuous.
Compare the existence of a continuous 1d parameterization $\varphi$ of a dense subset of all continuous functions with the following “negative” results:

**Is Hilbert’s Curve Injective or Surjective?**

- There is no continuous dense injection from $[0; 1] \rightarrow [0; 1]^2$ (because it would be a bijection)
- There is a continuous surjection $[0; 1] \rightarrow [0; 1]^2$ (space-filling curves)
- The $n$th approximation to Hilbert’s curve is *injective but not surjective* for all $n < \infty$.
- But Hilbert’s curve itself ($n \rightarrow \infty$) is *surjective but not injective*!
Why is all this important?

How can a strictly increasing $\sigma \in C^\infty$ be pathological?

One can actually even find entire real analytic $\sigma$.

The construction feels like cheating, but why is this cheating bad?

Ultimately we want/need to train NN and usually by (variants of) gradient descent.

Gradient descent produces a sequence of estimates $\Phi_k$ converging ideally to $f$ or an approximation thereof.

Implies $||\Phi_n - \Phi_m|| \to 0$ for $n, m \to \infty$ i.e. small change for large $n, m$.

A small change in $\Phi$ should be achievable by a small change in its parameters $W, b, c$.

Otherwise gradient descent has to travel arbitrarily far in parameter space, which likely does not work (well).
In the pathological stitching $\check{\sigma}$, moving from $\Phi_k$ to $\Phi_{k+1}$ requires jumping from one $\check{\sigma}$-cell $n_k$ ($\Phi_k = \varphi_{n_k} = \check{\sigma}(x \cdot +2n_k)$) to another far-away $\check{\sigma}$-cell $n_{k+1}$ ($\Phi_{k+1} = \varphi_{n_{k+1}} = \check{\sigma}(x \cdot +2n_{k+1})$), in-between even having to pass through bad approximations.

So a minimal reasonable requirement is that the parameters change continuously with $\Phi$.

This is stronger than $\Phi$ changing continuously with the parameters.

$\varphi : \mathbb{R}^d \to \mathcal{M}_r(\sigma)$ ($d = (n + 2)r$)
$\varphi : W, b, c \mapsto \Phi_{W,b,c}(\cdot)$ is continuous surjection.

If parameter symmetries are ignored, it is even a bijection.

Restrict parameter space so that $\varphi$ is injective, hence bijective

$\varphi^{-1} : \mathcal{M}_r(\check{\sigma}) \to \mathbb{R}^d$ is not continuous (similar argument as above)
## Table of Contents

1. Motivation/Preliminaries
2. Shallow Neural Networks
3. Universality = Density of 1HLP
4. Variations
5. Pathological Approximations
6. Degree of Continuous Approximation
7. Two Hidden Layer Perceptron (2HLP)
Continuous General Non-linear Approximation Lower Bound

- **Homeomorphism** between $\mathbb{R}^d$ and $C[0; 1]$ or dense subset thereof desirable but **not possible**.
- Find "**approximate homeomorphism**". Formally:
  - We want to approximate function $f \in B^m_p$  
  - $M_d : \mathbb{R}^d \rightarrow L^p$ any map from parameters $w$ to $M_d(w) = \Phi_w \approx f$ *(think: NN approximating function)*
  - Let $P_d : B^m_p \rightarrow \mathbb{R}^d$ be continuous *(intent: $P_d(f) = w$ is best approximation parameter)*
  - What is best $M_d$ and $P_d$ to approx. any $f \in B^m_p$ as $\Phi_w$ for some $w$?

**Theorem (Continuous general non-linear approximation lower bound)**

For $p \in [1; \infty]$, $m \geq 1$, $n \geq 1$, we have

$$\inf_{P_d, M_d} \sup_{f \in B^m_p} \| f - M_d(P_d(f)) \|_p \geq C d^{-m/n}$$
General Lower Bound Intuition

- **Intuition** for $m = 1$:

- Divide domain $B^n \subset [-1; 1]^n$ of $f$ into $(1/\varepsilon)^n$ grid cells.

- In order to describe an arbitrary 1-Lipschitz to accuracy $\varepsilon$, we need to record its e.g. average value in each cell.

- For $P_d$ to be continuous we need one real number per cell (parameter savings $\mathbb{R}^k \rightarrow \mathbb{R}$ would be discontinuous or lossy)

- Hence $d \geq (1/\varepsilon)^n$ is needed. Conversely $\varepsilon \geq d^{-1/n}$.

- Smoother functions require less fine grid ($\varepsilon \sim \varepsilon^{1/m}$)

- **Proof** uses Borsuk’s Antipodality Theorem. Maybe related to Hedgehog Theorem?

  You can’t comb a hedgehog flat
Corollary (Continuous Lower Bound for 1HLP)

For \( p \in [1; \infty], \ m \geq 1, \ n \geq 1, \) let \( Q_r : L^p \rightarrow M_r(\sigma) \) be any method of approximation where the parameters \( W, b, c \) depend continuously on the function \( f \) being approximated, or equivalently, \( Q_r \) is a continuous functional of \( f \), then \( \sup_{f \in B^m_p} \|f - Q_r(f)\|_{L^p(B^n)} \geq Cr^{-m/n} \).

Theorem (Non-Pathological Continuous Upper Bound for 1HLP)

For \( \sigma : \mathbb{R} \rightarrow \mathbb{R} \) such that \( \sigma \in C^\infty([a; b]) \setminus \text{Poly} \) for some \( a < b \) and any \( p \in [1; \infty], \ m \geq 1, \ n \geq 2, \) there is a bounded linear operator \( Q_r : L^p \rightarrow M_r(\sigma) \) such that for all \( f \in B^m_p, \|f - Q_rf\|_{L^p(B^n)} \leq Cr^{-m/n} \). In particular \( \inf_{\Phi \in M_r(\sigma)} \|f - \Phi\|_{L^p(B^n)} \leq Cr^{-m/n} \).

- Indeed, \( W \) and \( b \) can be chosen fixed independent of \( f \), and \( c \) depends linearly on \( f \).
- Bound valid for any smooth \( \sigma \) such as SIGMOID.
Proof Sketch

- As before, the vector space of polynomials $\mathcal{H}_k$ of exactly degree $k$ has dimension $s := \binom{n-1+k}{k} \approx k^{n-1}$ for $k \gg n$.

- As before, a linear combination of $s$ ridge functions based on 1d polynomials $\pi_k$ of degree at most $k$ can represent all multivariate polynomials $P_k \in \mathcal{P}_k$.

- $\pi_k \in \mathcal{N}_{k+1}(\sigma)$ i.e. representable by $k + 1$ neurons.

- Together this shows that $\mathcal{P}_k \subseteq \mathcal{M}_{(k+1)s}(\sigma)$ i.e. representable by $r := (k + 1)s \approx k^n$ neurons.

- Hence $\inf_{\Phi \in \mathcal{M}_r(\sigma)} \|f - \Phi\|_p = \inf_{\Phi \in \mathcal{M}_r(\sigma)} \|f - \Phi\|_p \leq \inf_{\Phi \in \mathcal{P}_k} \|f - \Phi\|_p \leq Ck^{-m} \approx Cr^{-m/n}$.

For analytic functions there are better-order approximations, again based on polynomials (Mhaskar, 1996).
Restricted Function Classes

The curse of dimensionality can only be overcome by considering restricted function classes. Generic Meta-Theorem:

**Theorem (Approximating Convex Combinations)**

- Let \( \varepsilon_r(K) := \min\{r : \text{ } r \text{ balls of radius } \varepsilon_r(K) \text{ can cover } K\} \).
- Let \( K \) be a bounded subset of a Hilbert space.
- Let \( f \) be in the convex hull of \( K \).
- Then there is a function \( f_r \) of the form \( f_r = \sum_{i=1}^{r} a_i g_i \)
  - with \( g_i \in K \) and \( a_i \geq 0 \) and \( \sum_{i=1}^{r} a_i \leq 1 \)
  - such that \( \|f - f_r\|_H \leq 2\varepsilon_r(K)/\sqrt{r} \).

**Trivial example:** For \( r = |K| < \infty \), we have \( \varepsilon_r(K) = 0 \) and \( f \) exact convex combination of all \( g_i \in K \).
Theorem (Approximating Functions with Nice Fourier Transform)

For functions $f$ with ‘nice’ Fourier transformation:

$$\inf_{\Phi \in \mathcal{M}_r(\sigma)} \| f - \Phi \|_p \leq C r^{-1/2}$$

- The formal definition of ‘nice’ is not nice
- Rate $r^{-1/2}$ is independent of dimension $n$
- **Intuition:** $\sin(k \cdot x)$ and $\cos(k \cdot x)$ in Fourier trafo are ridge functions, so easy to represent by linear combinations of ridge functions $\sigma(w \cdot x + b)$.
- **Solution** $\Phi$ can even be found iteratively by linearly mixing one new neuron at a time to an existing solution, keeping the old weights fixed, and only optimizing the new weights.
Two Hidden Layer Perceptron (2HLP)

- \( y = \Phi(x) := \sum_{k=1}^{s} d_k \sigma(\sum_{i=1}^{r} c_{ki} \sigma(\sum_{j=1}^{n} w_{ij} x_j + b_i) + a_k) \)
  \( \equiv d \cdot \sigma(C\sigma(Wx + b) + a) \)

- 2HLP is more powerful than 1HLP more powerful than 0HLP (in some ways).

- Little theoretical is known concerning (dis)advantages of more layers (compared to wider hidden layers)
In the $1HLP$, $\forall \sigma$, no $0 \neq g \in \mathcal{M}(\sigma)$ has compact support: 
$$\int_{\mathbb{R}^n} |g(x)|^p \, dx = \infty \text{ for } n > 1 \text{ and } p < \infty.$$ 

Proof: Ridge functions are const. in some direction, and $\int_{-\infty}^{\infty} c = \infty$.

This is no longer true in $2HLP$:

Choose $\sigma = \sigma_0 = [\cdot \geq 0] = \text{STEP}$, then 
$$\sigma_0(\sum_{i=1}^m \sigma_0(\mathbf{w}_i \cdot \mathbf{x} - b_i) + 1/2 - m) = \begin{cases} 1 \text{ if } \mathbf{w}_i \cdot \mathbf{x} \geq b_i \forall i \\
0 \text{ else.} \end{cases}$$

Can represent the characteristic function of any closed convex polygonal domain.

For example for $a_i < b_i$: Characteristic function of a hyper-cube 
$$\sigma_0(\sum_{i=1}^n (\sigma_0(x_i - a_i) + \sigma_0(-x_i + b_i)) - (2n - 1/2)) = [\mathbf{x} \in \prod_{i=1}^n]$$

$\sigma_0$ can be approximated by sigmoidal $\sigma(\lambda \cdot) \rightarrow \sigma_0$ for $\lambda \rightarrow \infty$.

$1HLP$ can approximate such compact functions on compacta, but only un-naturally and with many neurons.
Genuine Functions of 3 Variables

- For sure some functions of 2 variables are needed to create functions of $n$ variables by composition.
- Are there genuine functions of three variables? I.e. not (de)composable as functions of 1 and 2 variables.
- We can biject $\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}$ and hence recursively biject $\beta : \mathbb{R}^n \rightarrow \mathbb{R}$.
  With $\gamma : \mathbb{R} \rightarrow \mathbb{R}$ defined as $\gamma := f \circ \beta^{-1}$, then $f = \gamma \circ \beta$.
- With $x \equiv \sum_{i=-b}^{\infty} 2^{-b} x_i \in \mathbb{R}$, let $\delta(x) := \sum_{i=-b}^{\infty} 4^{-b} x_i \in \mathbb{R}$.
  Then $\alpha(x, y) := \delta(x + (y + y))$ is injection, hence
- All multivariate functions $f$ can be composed from univariate functions $\gamma$ and bivariate $+$.
- **Problem:** $\delta$ is totally discontinuous (very pathological)
- But key in *Boolean circuits* ($\mathbb{R} \rightsquigarrow \{0, 1\}$).
  Only $\text{OR} \hat{=}+$ and $\text{NOT} \hat{=} \gamma$ needed.
Kolmogorov Superposition Theorem

- Is it possible to \textit{exactly} represent any continuous multivariate function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as a combination of \textit{continuous} univariate functions $\gamma_i : \mathbb{R} \rightarrow \mathbb{R}$ and the single binary function ‘$+$’?
- Seems hopeless, but ...

\textbf{Theorem (Improved Kolmogorov Superposition Theorem)}

There exist $n$ constants $\lambda_j > 0$, $\sum_{j=1}^n \lambda_j \leq 1$, and $2n + 1$ strictly increasing continuous functions $\phi_i : [0; 1] \rightarrow [0; 1]$, all independent of $f$, such that every continuous function $f : [0; 1]^n \rightarrow \mathbb{R}$ can be represented in the form

$$f(x_1, ..., x_n) = \sum_{i=1}^{2n+1} g(\lambda_1 \phi_i(x_1) + ... + \lambda_n \phi_i(x_n))$$

for some continuous $g[0; 1] \rightarrow \mathbb{R}$ depending on $f$.

- The $\phi_i$ are based on Cantor functions = Devil’s staircase, which are even more pathological than $\tilde{\sigma}$.
- \textit{Proof:} Whole PhD theses have been devoted, e.g. [Act18].
Even allowing pathological $\bar{\sigma}$, there was an intrinsic lower bound on the degree of approximation achievable with 1HLP depending on the number of neurons used.

Not so for 2HLP:

**Theorem (Universality of pathological bounded-size 2HLP)**

A 2HLP with $\sigma = \bar{\sigma}$ and $(4n + 3)(2n + 1)$ resp. $4n + 3$ hidden neurons in the first (second) layer can uniformly approximate any continuous function to arbitrary precision.
Choose $\phi_i$ and $g$ in Kolmogorov’s Sup. Thm. to represent $f$.

Approximate $\phi_i$ ($g$) by the first (second) layer in 2HLP.

$g, \phi_i \in \overline{N_1(\ddot{\sigma})}$, i.e. each approximable by one $\ddot{\sigma}$-neuron.

Hence we need $n(2n + 1)$ resp. $2n + 1$ neurons in first (second) layer.

If we want $\ddot{\sigma}$ to be monotone increasing, we need 3 neurons each.

The 2 extra neurons linearly slant functions to (de)monotonize them.

By combining linear neurons we only need $(4n + 3)(2n + 1) + (4n + 3)$ overall.
More Pathological Results

- Recurrent NN with $\sigma=$HARD-TANH and integer/rational/real weights can compute any regular/recursive/arbitrary partial functions in linear/linear/exponential time [SS92].

- There exist recurrent NN with 1000 neurons which can simulate a Universal TM [SS92]. Proof idea: 2-stack FSM is Turing complete. Store stack in bits of real number.

- Recurrent NN can even do hyper-computation and represent any function [SS94].

- Improved rates for Deep NN with ReLU $\sigma$ by tiling input and predicting bits of real output [LSYZ20].
Summary

- (Non)Asymptotic approximation results mostly for 1HLP
- Surprisingly few neurons are needed for exact interpolation
- 0HLP too limited. 2HLP have some extra advantages
- Important to distinguish pathological from genuine results
- E.g. parameters should change gradually with the target function
- Approximation is necessary but not sufficient for learning
- Most activation functions are ok (in theory as well as practice)
- No way out of curse of dimensionality unless restricting function class
- Smooth functions require fewer neurons to approximate
- Proof tools: Weierstrass approx., Ridge functions, reduction to 1d
- NN approximation theory is just the beginning ...
References

Jonas Actor.

Mikhail Belkin.
*Fit without Fear: An Interpolation Perspective on Generalization and Optimization in Modern Machine Learning*, November 2018.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White.


Allan Pinkus.