Nonlinear Approximation by Deep ReLU Neural Networks

Ronald DeVore

Collaborators: Ingrid Daubechies- Simon Foucart- Boris Hanin - Guergana Petrova
Deep Neural Networks

Deep Neural Networks have had remarkable success in targeted applications - especially Learning.

However, even its proponents agree that there are insufficient foundational results to explain this success.

The success of Deep Neural Networks must at least partially lie in increased ability to approximate functions vis a vis the more traditional methods.

This talk will examine the approximation properties of Deep Neural Networks in the simplest setting.

- ReLU activation
- approximation of univariate functions

We want to provide Mathematical Backbone for DL that will guide us as to when and how to apply these methods and to certify (guarantee) their performance.
ReLUs Networks

- Recall that for an input vector \((x_1, \ldots, x_d)\)
  \[
  \text{ReLU}(x_1, \ldots, x_d) = (\text{ReLU}(x_1), \ldots, \text{ReLU}(x_d))
  \]
  where \(\text{ReLU}(y) := y_+ := \max\{y, 0\}, \ y \in \mathbb{R}\)

- We consider ReLUs networks with width \(W\) and depth \(L\)
  and full connectivity

- \(\Upsilon^{W,L}\) denotes the set of outputs of such a network

- The network produces a Continuous Piecewise Linear Function (CPwL) \(\eta_{i,L}\) at the node \(i, L\)

- **First Layer:** for input \(x \in \mathbb{R}\), \(\eta_{i,1}(x) = (a_i x + b_i)_+\), \(i = 1, \ldots, W\), with \(a_i, b_i \in \mathbb{R}\)

- **Layers** \(L \geq 2\), \(\eta_{i,L} = \sum_{j=1}^{W} [a_{i,j} \eta_{j,L-1} + b_i]_+\)
Architecture of Neural Networks

\[
\begin{align*}
    &\mathbb{R} \xrightarrow{\psi} x^{(0)} \\
    &\mathbb{R}^3 \xrightarrow{\psi} x^{(1)} \\
    &\mathbb{R}^3 \xrightarrow{\psi} x^{(2)} \\
    &\mathbb{R}^3 \xrightarrow{\psi} x^{(L-1)} \\
    &\mathbb{R}^3 \xrightarrow{\psi} x^{(L)} \\
    &\mathbb{R} \xrightarrow{\psi} x^{(L+1)}
\end{align*}
\]
Structure of $\Upsilon^{W,L}$

- Each $\eta_{i,j}$ and the output $S$ of the network is a CPwL
- $\Upsilon^{W,L}$ is a nonlinear space depending on $n = n(W, L) \approx W^2L$ parameters
- $\Upsilon^{W,1}$ consists of all CPwL functions with $W$ break points
  - This space is usually denoted by $\Sigma_W$ and is well understood - even as we allow $W$ to get large: the parameters are the breakpoints and slopes

However, we are interested in the other direction when $W$ is fixed and $L$ gets large: The effect of depth

In this case, for short, we denote by $\Upsilon_n$, $n = 1, 2, \ldots$ the largest of the spaces $\Upsilon^{W,L}$ with at most $n$ parameters

This is a nested family $\Upsilon_n \subset \Upsilon_{n+1}$, $n \geq 1$
Comparing $\Upsilon_n$ with $\Sigma_n$

- Both of these are nonlinear spaces of CPwL functions depending on $\sim n$ parameters
- Does $\Upsilon_n$ have approximation power that $\Sigma_n$ does not have?

First Theorem (DDFHP) $\Sigma_n \subset \Upsilon_{12n}, n \geq W$

Architecture used in Proof
How do we measure performance?

- Does $\Upsilon_n$ have approximation power that $\Sigma_n$ does not have?

- Given a Banach space $X$ with norm $\| \cdot \|_X$ and $F \in X$
  
  $\sigma_n(F)_X := \inf_{S \in \Sigma_n} \| F - S \|_X$
  $\nu_n(F)_X := \inf_{S \in \Upsilon_n} \| F - S \|_X$

- For $n \geq 1$ we have $\nu_{12n}(F)_X \leq \sigma_n(F)_X$

- So approximation by deep networks is at least as powerful as their shallow cousins

- How would we show deep networks are far more powerful

  - Show model classes $K$ of functions that are well approximated by $(\Upsilon_n)_{n \geq 1}$ but they are not well approximated by $(\Sigma_n)$ or any other classical approximation methods: Fourier, wavelets, adaptive gridding, etc.
Approximation Classes

There is a second standard way to understand the approximation power of a particular method of approximation.

We try to understand (characterize) the functions that are approximated with a certain error rate.

We discuss only approximation in the uniform norm, i.e. $X = C[0,1]$ and introduce approximation classes.

Given an $r > 0$ we define the approximation class $A^r(\Upsilon)$ as the set of all $f \in C[0,1]$ for which

$$\nu_n(F) \leq Mn^{-r}, \quad n \geq 1$$

The smallest $M$ is by definition $|f|_{A^r(\Upsilon)}$.

We define the classes $A^r(\Sigma)$ in a similar way.
The structure of $\Upsilon_n$ and $\Sigma_n$

- Goal is to show that $A^r((\Upsilon_n))$ is a lot larger than $A^r((\Sigma_n))$
- By the way $A^r((\Sigma_n))$, $0 < r < 2$ is precisely known (Besov spaces)
- To do this we first show $\Upsilon_n$ is a lot larger than $\Sigma_n$
- What properties does $\Upsilon_n$ have that $\Sigma_n$ does not?
- If $S \in \Upsilon_n$, $T \in \Upsilon_m$ then $S + T \in \Upsilon_{n+m}$
  - But $\Sigma_n$ has the same addition property
- If $S \in \Upsilon_n$, $T \in \Upsilon_m$ then $S \circ T := S(T) \in \Upsilon_{n+m}$
  - $\Sigma_n$ does not have this property
  - The best we can say in this case is that $S \circ T \in \Sigma_{nm}$
- Everyone knows this composition property is somehow behind the success of depth
Example: Self Similar Functions in $\Upsilon_n$

- Let $H$ be the hat function on $[0, 1]$
- The $n$-fold composition $H^n$ is in $\Upsilon_n$ but not in $\Sigma_k$ unless $k \geq 2^n$
More general construction

There is a more general construction

Let $S \in \mathcal{Y}_k$ and vanish at $0, 1$
Let $J_1, \ldots, J_m$ be $m$ disjoint intervals
Then $F(x) = \sum_{i=1}^{m} S(h_i(x - a_i))$, is in $\mathcal{Y}_{2(k+m)}$

We can do higher order composition
Consequences

- We draw some consequences from these results

First note the following Simple General Principle

- Let $f_k \in \Upsilon_k$, $k = 1, 2, \ldots$ with $\|f_k\| = 1$
- Let $(\beta_k)_{k \geq 1} \in \ell_1$
- Then $F := \sum_{k \geq 1} \beta_k f_k$ satisfies

$$v_n^2(F)_{C[0,1]} \leq \sum_{k > n} |\beta_k|, \quad n \geq 1$$

- When the $f_k = \phi \circ k$ for a fixed function $\phi \in \Upsilon_m$, then $n^2$ can be changed to $mn$.
- When we fix one $\phi$ we can view $(\phi^\circ n$ as a dictionary
- $\mathcal{F}(\phi) := \{ F : F = \sum_{n \geq 1} c_n \phi^\circ n \}$ is the span of this dictionary
Extremes

Let us note some functions $F$ that can be approximated to **exponential accuracy** by deep nets, i.e.,

$$v_n(F) \leq Cr^n, \ n \geq 1 \text{ and } r \in (0, 1)$$

**Analytic Functions:**

$$F(x) = \sum_{k \geq 0} c_k x^k; \ (c_k)_{k \geq 1} \in \ell_1$$

A lot of people have observed this:

$$x(1 - x) = \sum_{n=1}^{\infty} 4^{-n} H \circ n$$

**The Tagaki functions:**

$$F = \sum_{k \geq 1} t^k g(\psi \circ k), \quad |t| < 1,$$

with $g, \psi \in \mathcal{Y}_k$ and $\psi : [0, 1] \to [0, 1]$

**The Weierstrass function:**

$$F = \sum_{k=1}^{\infty} 2^{-k} H \circ k$$

which is nowhere differentiable

So all smooth functions and some very rough functions in $\mathcal{A}^r((\mathcal{Y}_n))$ for all $r > 0$ none of them are in $\mathcal{A}^r((\Sigma_n))$, if $r$ is large enough
One common theme is to show \( \mathcal{Y}_n \) approximates as well as nonlinear \( n \) term approximation from common bases-frames-dictionaries: e.g. Fourier, wavelets, etc.

Consider the following sinusoidal like Riesz basis

\[
C = \begin{pmatrix} 1 & -1 & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ -1 & -1 & -1 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & -1 & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ -1 & -1 & -1 \end{pmatrix}, \quad C_3 = \begin{pmatrix} 1 & -1 & 1 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ -1 & -1 & -1 \end{pmatrix}, \quad S_3 = \begin{pmatrix} 1 & -1 & 1 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ -1 & -1 & -1 \end{pmatrix}
\]
This basis

- This family of functions is a Riesz basis for $L_2[0, 1]$.
- Each of the functions $S_j, C_j, j = 1, \ldots, 2^n$, is in $\Upsilon_{cn}$ with $c$ an absolute constant.
- Any sum of $k$ terms from the first $2^n$ elements is in $\Upsilon_{ckn}$ represented by $ckn$ parameters.
- These results are similar to compressed sensing sparse recovery.
- Here something interesting happens: Deep nets (on the surface) seem to handle $k$ term approximation from exponentially large dictionary - more on this later.
Other Fantastic Results

- One can derive from the self similar theorem the following result of Yarotski

If $F \in \text{Lip} \ 1$ then

$$
\nu_n(F) \leq \frac{C\|F\|_{\text{Lip} \ 1}}{n \log n}, \quad n \geq 1
$$

- A second disturbing result (not for ReLU activation) is that of Maiorov and Pinkus

  This result holds for functions in $C[0, 1]^d$ but I take only $d = 1$

- **MP Theorem:** There exists an activation function $\sigma$ and a neural network with $W = 9$ and $L = 3$ such that the outputs of this network are dense in $C[0, 1]$

- The outputs are a space filling manifold!
Let us be careful

- The above results say there is an approximating function in $\Upsilon_n$ with the advertised performance.
- It does not tell us how to find such an approximation given $F$ or some data about $F$.
- $\Upsilon_n$ is a parametric manifold.
  - If $y \in \mathbb{R}^n$ are the parameters of the NN then let $M(y)$ denote its output which is a CPwL.
- To build an approximation scheme, we need to construct a mapping $a : C[0, 1] \mapsto \mathbb{R}^n$.
  - $A(F) := M(a(F))$ is the approximation to $F$.
- Approximation procedures of this type are called Manifold Approximation.
Manifold Approximation

So one way to understand if NNs offer some unexpected and unheard of performance is to compare it with other approximation methods (not necessarily Neural Networks) of the same form.

Let $A_n := \{(a,M) : a : X \to \mathbb{R}^n, M : \mathbb{R}^n \to X\}$.

The error in approximating $F$ by the pair $(a,M)$ is

$$E_{a,M}(F) := \|(F - A(F))\|_X$$

The error in approximating a class $K$ is

$$E_{a,M}(K) := \sup_{F \in K} \|(F - A(F))\|_X$$
Manifold Approximation

In evaluating manifold approximation a critical issue is what conditions are imposed on the mappings \( a, M \).

We describe three possibilities and their corresponding widths measuring best possible performance:

- No conditions: \( \delta_n^0(K) := \inf_{(a,M) \in A_n} E_{a,M}(K)X \)
- Continuous \( \delta_n(K) := \inf_{(a,M) \in A_n, \text{cont}} E_{a,M}(K)X \)
- Lipschitz \( \delta_n^*(K) := \inf_{(a,M) \in A_n} \text{Lipschitz } E_{a,M}(K)X \)
Three Theorems

Theorem 1: \( \delta_1^0(K)_X = 0 \) for any compact set \( K \subset X \)
Space filling curves

Theorem 2: For \( K \) the unit ball in \( \text{Lip} \, 1 \):
\[ \delta_n(K)_X \approx n^{-1} \]
DeVore-Kyriazis-Leviaton-Tikhomirov
Similar results for Besov balls

Theorem 3: Cohen-DeVore-Petrova-Wojtaszczyk
If \( X \) is any Banach space and \( K \subset X \) is compact then
\[ d_n^*(K) = O(n^{-r}) \text{ implies } \epsilon_n(K)_X = O(n^{-r}) \]
If \( X \) is a Hilbert space and \( K \subset X \) is compact
\[ \delta_{32n}^*(K)_X \leq C \epsilon_n(K)_X \]
Here \( \epsilon_n(K)_X \) are the entropy numbers of \( K \): The smallest \( \epsilon \) for which we can cover \( K \) by \( 2^n \) balls of radius \( \epsilon \)
Last Thoughts

- Can we precisely characterize $A^r(\mathcal{T}_n)$? Maybe

- Can we construct natural classes $K$ for which $\nu_n(K)_X$ tends to zero slower than $\epsilon_n(K)_X$ We have already done this. We probably can do more in this direction.

- At least part of the success of deep networks stems from their space filling.
  - Stability is always an issue with space filling curves and manifolds: We need to sort out which functions are captured stably.

- For learning problems we only have access to $F$ from data samples.

- How does this fact of partial information effect approximation rates/stable performance.