Does Size Matter?

Universal Approximation and the Efficiency of Depth

Oscar I. Hernandez

Department of Mathematics & Statistics
University of Alaska Fairbanks

March 31, 2022

Goal

Given:

1. $p : \mathbb{R}^n \to \mathbb{R}$ a multivariate polynomial of degree $d \in \mathbb{N}$
2. $\sigma : \mathbb{R} \to \mathbb{R}$ in C^d with $x_0 \in \mathbb{R}$ satisfying $\left[\frac{d^r\sigma}{dx^r}\right]_{x_0} \neq 0$ for all $r \leq d$
3. open box $(-R, R)^n \subset \mathbb{R}^n$ for some $R \geq 0$

Theorem (Rolnick and Tegmark [2017])

Let $m_k^\varepsilon(p)$ be the minimum of neurons in a depth-k network N satisfying $||N - p||_\infty < \varepsilon$ on $(-R, R)^n$. If $d > 1$, then

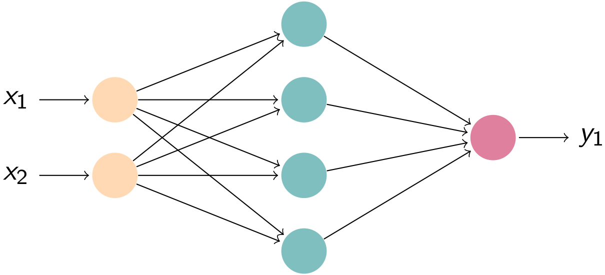
$$\lim_{\varepsilon \to 0} m_d^\varepsilon(p) < \lim_{\varepsilon \to 0} m_1^\varepsilon(p) < \infty.$$

Strategy

1) Approximate $p_2(u, v) = uv$. 2) Replicate proof technique. 3) $$$

Outline

Problem Architecture



Theorem (Lin et al. [2017])

Given the bivariate monomial $p_2(x_1, x_2) = x_1 x_2$ *and a tolerance* $\varepsilon > 0$, *there is a shallow neural network* N *with 2 inputs, m hidden neurons, and 1 output such that* $||N - p_2||_\infty < \varepsilon$ *on* $(-R, R)^2$. *This requires* $m = 4$ *exactly.*

Strategy

1) Construct N *such that* $||N - p_2||_\infty < \varepsilon$ *on* $(-\varepsilon, \varepsilon)^2$. *2) Scale. 3) $$$*

Solution Construction: First Affine Transformation $A^{[1]}$

$$
\begin{aligned}
A^{[1]}\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) &= W^{[1]}\begin{bmatrix} u \\ v \end{bmatrix} + b^{[1]} \\
&= \begin{bmatrix} +1 & +1 \\ -1 & -1 \\ +1 & -1 \\ -1 & +1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} +u + v + 0 \\ -u - v + 0 \\ +u - v + 0 \\ -u + v + 0 \end{bmatrix}
\end{aligned}
$$

Solution Construction: Activation $\vec{\sigma} \circ A^{[1]}$

$$
(\vec{\sigma} \circ A^{[1]}) \left(\begin{bmatrix} u \\ v \end{bmatrix} \right) = \vec{\sigma} \left(\begin{bmatrix} +1 & +1 \\ -1 & -1 \\ +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right)
$$

$$
= \begin{bmatrix} \sigma(+u+v) \\ \sigma(-u-v) \\ \sigma(+u-v) \\ \sigma(-u+v) \end{bmatrix}
$$

Solution Construction: $f = A^{[2]} \circ \vec{\sigma} \circ A^{[1]}$

$$f\left(\begin{bmatrix} u \\ v \end{bmatrix}\right) = \frac{1}{4\sigma_2}\begin{bmatrix} +1 & +1 & -1 & -1 \end{bmatrix} \vec{\sigma}\left(\begin{bmatrix} +1 & +1 \\ -1 & -1 \\ +1 & -1 \\ -1 & +1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) + \begin{bmatrix} 0 \end{bmatrix}$$

$$= \frac{\sigma(+u+v) + \sigma(-u-v) - \sigma(+u-v) - \sigma(-u+v)}{4\sigma_2}$$

Lemma

Then, f quartically approximates p as follows.

$$|f(u,v) - uv| \in o(u^2 + v^2)uv$$

For any $\varepsilon > 0$ in particular, if $|u|, |v| < \sqrt[4]{\varepsilon/2}$ then $|f(u,v) - uv| < \varepsilon$.

Proof of Proposition

Proof.

Let $m(u,v) = f\left(\begin{bmatrix} u \\ v \end{bmatrix}\right)$. By Taylor's theorem, $\exists \{\xi_k\}_{k=1}^{2^2}$ such that

$4m(u,v)\sigma_2 = \sigma(u+v) + \sigma(-u-v) - \sigma(+u-v) - \sigma(-u+v) =$

$$\begin{array}{cccc}
+\frac{\sigma_0}{1}(+u+v)^0 & +\frac{\sigma_0}{1}(-u-v)^0 & -\frac{\sigma_0}{1}(+u-v)^0 & -\frac{\sigma_0}{1}(-u+v)^0 \\
+\frac{\sigma_1}{1}(+u+v)^1 & +\frac{\sigma_1}{1}(-u-v)^1 & -\frac{\sigma_1}{1}(+u-v)^1 & -\frac{\sigma_1}{1}(-u+v)^1 \\
+\frac{\sigma_2}{2}(+u+v)^2 & +\frac{\sigma_2}{2}(-u-v)^2 & -\frac{\sigma_2}{2}(+u-v)^2 & -\frac{\sigma_2}{2}(-u+v)^2 \\
+\frac{\sigma_3}{6}(+u+v)^3 & +\frac{\sigma_3}{6}(-u-v)^3 & -\frac{\sigma_3}{6}(+u-v)^3 & -\frac{\sigma_3}{6}(-u+v)^3 \\
+\frac{\sigma_4}{24}(+u+v)^4 & +\frac{\sigma_4}{24}(-u-v)^4 & -\frac{\sigma_4}{24}(+u-v)^4 & -\frac{\sigma_4}{24}(-u+v)^4 \\
+\frac{\sigma^{(5)}(\xi_1)}{120}(+u+v)^5 & +\frac{\sigma^{(5)}(\xi_2)}{120}(-u-v)^5 & -\frac{\sigma^{(5)}(\xi_3)}{120}(+u-v)^5 & -\frac{\sigma^{(5)}(\xi_4)}{120}(-u+v)^5
\end{array}$$

$$\begin{aligned}
m(u,v) &= \frac{1}{4\sigma_2}\left[0 + \frac{0}{1} + \frac{\sigma_2}{2}(8uv) + \frac{0}{6} + \frac{\sigma_4}{24}(16u^3v + 16uv^3) + \frac{4}{120}o\left((u+v)^5\right)\right] \\
&= 0 + \frac{4\sigma_2}{4\sigma_2}(uv) + \frac{(u^2+v^2)\sigma_4}{6\sigma_2}(uv) + \frac{o\left((u+v)^4\right)}{30\sigma_2} \\
&= uv\left[1 + o(u^2+v^2)\right] \to uv \text{ as } |u|, |v| \to 0
\end{aligned}$$

Solution Construction: Scale

Let $R > 0$, $\varepsilon > 0$, and set $\lambda = \frac{\varepsilon/2}{\max(R,1)}$.
Given $x_1, x_2 \in (-R, R)$, let $u = \lambda x_1$, $v = \lambda x_2$ so that $u, v \in (-\varepsilon, \varepsilon)$.

$$\begin{aligned}
N\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) &= f\left(\lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)/\lambda^2 \\
&= \frac{\lambda^{-2}}{4\sigma_2}\begin{bmatrix} +1 & +1 & -1 & -1 \end{bmatrix} \vec{\sigma}\left(\lambda \begin{bmatrix} +1 & +1 \\ -1 & -1 \\ +1 & -1 \\ -1 & +1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) + \begin{bmatrix} 0 \end{bmatrix} \\
&= \frac{\sigma(+u+v) + \sigma(-u-v) - \sigma(+u-v) - \sigma(-u+v)}{4\lambda^2\sigma_2} \\
&= m(u,v)/\lambda^2
\end{aligned}$$

Solution Construction: Proof

Theorem (Lin et al. [2017])

Given $R > 0$ *and* $\varepsilon > 0$*, there is a shallow neural net* N *with* $m = 2^2$ *hidden neurons satisfying* $|N(x_1, x_2) - x_1 x_2| < \varepsilon$ *for all* $(x_1, x_2) \in (-R, R)^2$*.*

Proof.

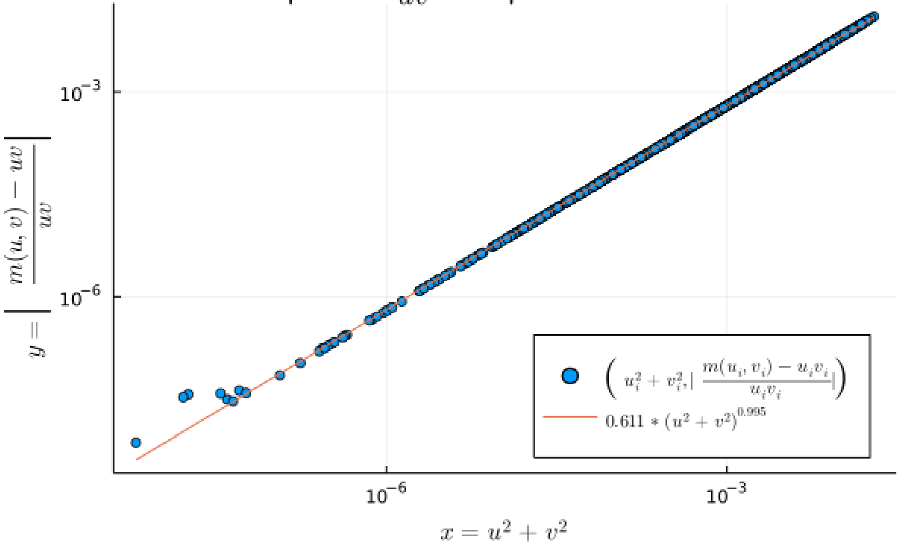
Set $\lambda = \frac{\varepsilon/2}{\max(R,1)}$.
Given $x_1, x_2 \in (-R, R)$, let $u = \lambda x_1$ and $v = \lambda x_2$ so that $u, v \in (-\varepsilon, \varepsilon)$.

$$N(x_1, x_2) = m(u, v)/\lambda^2 \to \frac{u}{\lambda}\frac{v}{\lambda} = x_1 x_2.$$

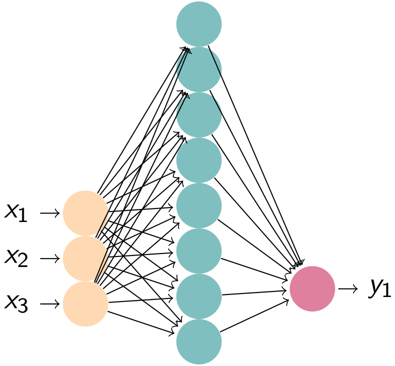
□

Numerical Analysis

$$\left| \frac{m(u,v) - uv}{uv} \right| \in o(u^2 + v^2)$$



k-ary Multiplication



Theorem (Lin et al. [2017])

Given the multivariate monomial $p_n(x) = \prod_{i=1}^{n} x_i$ *and a tolerance* $\varepsilon > 0$*, there is a shallow neural network* N *with* n *inputs,* m *hidden neurons, and* 1 *output such that* $||N - p_n||_\infty < \varepsilon$ *on* $(-R, R)^n$*.*
This requires $m = 2^n$ *exactly.*

Monomial of degree k

Corollary (Lin et al. [2017])

Given the multivariate monomial $p_n(x) = a\prod_{i=1}^{n} x_i$ *and a tolerance* $\varepsilon > 0$*, there is a shallow neural network* N *with* n *inputs,* m *hidden neurons, and* 1 *output such that* $||N - p_n||_\infty < \varepsilon$ *on* $(-R, R)^n$*.*
This requires $m = 2^n$ *exactly.*

Proof.

Let $N = A^{[2]} \circ \vec{\sigma} \circ A^{[1]}$ as above and let $A_a^{[2]} = aA^{[2]}$.

$$N_a = A_a^{[2]} \circ \vec{\sigma} \circ A^{[1]} = a \cdot N \to a\prod_{i=1}^{n} x_i$$

□

Polynomial of degree k

Theorem (Lin et al. [2017])

Given the multivariate polynomial $p(x) = \sum_{i=1}^{n} p_i(x)$ *and a tolerance* $\varepsilon > 0$, *there is a shallow neural network* N *with* *n inputs*, *m hidden neurons*, *and* *1 output* *such that* $||N - p_n||_\infty < \varepsilon$ *on* $(-R, R)^n$.

Proof.

Approximate monomial $p_i(\vec{x}) = a_i \prod_{j=1}^{n} x_j^{n_j}$ with $N_i = A_i^{[2]} \circ \vec{\sigma} \circ A_i^{[1]}$.

Let $A^{[1,2]} = \sum_{i}^{n} A_i^{[1,2]}$.

$$N = A^{[2]} \circ \vec{\sigma} \circ A^{[1]} = \sum_{i=1}^{n} N_i \to \sum_{i=1}^{n} p_i(\vec{x}) = p(\vec{x})$$

Universal Approximation Theorem

Theorem (Cybenko [1989])

Given the continuous function $f : \mathbb{R}^n \to \mathbb{R}$ and a tolerance $\varepsilon > 0$, there is a shallow neural network N with n inputs, m hidden neurons, and 1 output such that $||N - f||_\infty < \varepsilon$ on any $(-R, R)^n$. [a]

[a] Original theorem about any compact $K \subset \mathbb{R}^n$ follows from this.

Proof ($\varepsilon/2$-argument via [Lin et al., 2017]).

Pick p_d such that $||p_d - f||_\infty < \varepsilon/2$ on $[-R, R]^n$ via Stone-Weierstrass.
Pick N such that $||N - p_d||_\infty < \varepsilon/2$ on $(-R, R)^n$.

$$||N - f||_\infty \leq ||N - p_d||_\infty + ||p_d - f|| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

It is clear from the construction of N that $\lim_{\varepsilon \to 0} m_1^\varepsilon(p_d) < \infty$. □

Superior to [Cybenko, 1989] for which m grows as ε shrinks.

Asymptotic Depth Case

Theorem (Rolnick and Tegmark [2017])

$$\lim_{\varepsilon \to 0} m_k^{\varepsilon} \left(\prod_{i=1}^{n} x_i \right) = \mathcal{O}\left(n^{(k-1)/n} \cdot 2^{n^{1/k}} \right)$$

References

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6): 1223–1247, 2017.

David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. *arXiv preprint arXiv:1705.05502*, 2017.

Shallow Artificial Neural Network: Definition

Given $W = (w_{ij}) : \mathbb{R}^n \to \mathbb{R}^m$ and $b = (b_i) \in \mathbb{R}^m$, define $A : \mathbb{R}^n \to \mathbb{R}^m$ by $Ax = Wx + b$. Example:

$$A : \begin{bmatrix} u \\ v \end{bmatrix} \mapsto \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \\ w_{41} & w_{42} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} w_{11}u + w_{12}v + b_1 \\ w_{21}u + w_{22}v + b_2 \\ w_{31}u + w_{32}v + b_3 \\ w_{41}u + w_{42}v + b_4 \end{bmatrix}$$

Given $\sigma : \mathbb{R} \to \mathbb{R}$, define $\vec{\sigma} : \mathbb{R}^m \to \mathbb{R}^m$ by $(\vec{\sigma}(x))_i = \sigma(x_i)$

A "hidden" layer with m neurons is a composition $\vec{\sigma} \circ A : \mathbb{R}^n \to \mathbb{R}^m$.

A depth-k neural network is the pre-composition of A_{k+1} with k layers.

A shallow neural network is a depth-1 neural network.

Real k-ary Multiplication

1. Enumerate $\{S_j\}_{j=1}^{2^k} = 2^{[k]}$ and let $a_{ij} = s_i(S_j) = 2\left(1 - \chi_{S_j}(i)\right) - 1$
2. Let $w_j = \frac{1}{2^k n! \sigma_n} \prod_{i=1}^{n} a_{ij} = \frac{(-1)^{|S_j|}}{2^n n! \sigma_n}$ and $f = \sum_{j=1}^{2^m} w_j \vec{\sigma} \left(\sum_{i=1}^{n} a_{ij} x_i \right)$
3. If $p(x)$ lacks x_1 then terms in Taylor expansion cancel.
4. If $p(x) = \prod_{i=1}^{n} x_i$ then coefficients add to 1.