Does Size Matter?

On Neural Expressivity and Complexity

Oscar I. Hernandez

Department of Mathematics & Statistics
University of Alaska Fairbanks

March 29, 2022

Goal

Given:

1. $d, k \in \mathbb{N}$ and $R \in \mathbb{R}$
2. $p : (-R, R)^n \to \mathbb{R}$ a multivariate polynomial of degree d
3. $\sigma : \mathbb{R} \to \mathbb{R}$ in C^d such that $\exists x_0 \in \mathbb{R}$ satisfying $\forall r \leq d$, $\left[\frac{d^r\sigma}{dx^r}\right]_{x_0} \neq 0$.

Theorem (Rolnick and Tegmark [2017])

Let $m_k^\varepsilon(p)$ be the minimum of neurons in a depth-k network $N : \mathbb{R}^n \to \mathbb{R}$ satisfying $\sup\limits_{x\in(-R,R)^n} |N(x) - p| < \varepsilon$. Then, $\lim\limits_{\varepsilon\to 0} m_k^\varepsilon(p) < \infty$.

Strategy

Show result for $k = 1$, i.e. for shallow artificial neural networks.

Outline

Shallow Artificial Neural Network

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.

Example: Continuous 2-ary multiplication gate

Given non-linear $\sigma \in C^2$ with $\sigma_r = \sigma^{(r)}(0) \neq 0$ for $r \leq 2$, $u', v' \in \mathbb{R}$, let $\lambda = \frac{1/3}{\max(|u'|,|v'|,1)}$. Let $u = \lambda u'$, $v = \lambda v'$ so that $|u| + |v| < 1$. Consider:

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

Let $B(r) = B_{|r/2|}(r/2)$. For any x, there is an $\xi \in B(x)$ satisfying:

$$
\sigma(x) = \left(\sum_{k=0}^{4} \frac{\sigma^{(k)}(0)}{k!} (x-0)^k \right) + \frac{\sigma^{(5)}(0)}{5!} (\xi)^5.
$$

Real 2-ary Multiplication

$4m(u,v)\sigma_2 =$

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

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

Continuous multiplication gate

Theorem (Lin et al. [2017])

Can approximate multiplication with a single hidden layer consisting of 2^2 *neurons.*

Proof.

$$f(u', v') = m(u, v)/\lambda^2 \to \frac{u}{\lambda}\frac{v}{\lambda} = u'v'$$

$\square$

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.

Monomial of degree k

Scale final affine transformation by coefficient.

Polynomial of degree k

Approximate each monomial and add.

Universal Approximation Theorem

Superior to [Cybenko, 1989] for which m grows as ε shrinks.
We apply something like the Stone-Weierstrass theorem to extend result to continuous functions.

Depth

Can drop exponential number of neurons to linear with depth.

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.