

Approximation by Superpositions of a Sigmoidal Function (Cybenko, 1989)

Satvik Saha (ss7228)

October 28, 2025

The flexibility of a neural network lies in its ability to approximate a very large class of functions well, given an appropriate choice of activation function σ ; results of this nature are often termed as *universal approximation theorems*. Here, we will only look at one-hidden-layer networks $\Omega: I \rightarrow \mathbb{R}$ with arbitrary width where $I := [0, 1]^n$. These are maps of the form

$$x \mapsto \sum_{j=1}^N \alpha_j \sigma(y_j^\top x + \theta_j);$$

call this class of functions $\mathcal{S}_\sigma$. Cybenko [1989] shows that for continuous activation functions σ satisfying $\lim_{t \rightarrow -\infty} \sigma(t) = 0$ and $\lim_{t \rightarrow \infty} \sigma(t) = 1$, called *sigmoidal* activation functions, the class $\mathcal{S}_\sigma$ is dense in $(C(I), \|\cdot\|_\infty)$. This is also true of continuous σ satisfying a more general *discriminatory* property: given a finite signed regular measure μ on I ,

$$\int \sigma(y^\top x + \theta) d\mu(x) = 0 \text{ for all } y \in \mathbb{R}^n, \theta \in \mathbb{R} \implies \mu = 0.$$

The proof of the result in Cybenko [1989] is as follows: if $\mathcal{R} := \overline{\mathcal{S}_\sigma} \neq C(I)$, then one can produce a non-zero bounded linear functional L such that $L|_{\mathcal{R}} = 0$ via Hahn-Banach extension. Riesz Representation gives a finite signed regular measure μ_L such that $L = \int \cdot d\mu_L$. Consider the family of maps $\mathcal{H} := \{h_{y,\theta} : y \in \mathbb{R}^n, \theta \in \mathbb{R}\}$ where each $h_{y,\theta}(x) = \sigma(y^\top x + \theta)$; since $\mathcal{H} \subset \mathcal{S}_\sigma \subset \mathcal{R}$, we have $L|_{\mathcal{H}} = 0$. The discriminatory nature of σ forces $\mu_L = 0$, contradicting $L \neq 0$.

For measurable functions f on I in general, given $\epsilon > 0$, Lusin's theorem gives $f|_D = g$ continuous for some $D \subset I$ with Lebesgue measure $m(D) > 1 - \epsilon$. In this sense, neural networks can well approximate this broader class of measurable functions.

This result does not hold if σ is a polynomial. Indeed, $h_{y,\theta}$ would also be a polynomial, and $\int h_{y,\theta} d\mu = 0$ can only guarantee that finitely many moments of μ are zero, hence σ is not discriminatory. More directly, $\mathcal{S}_\sigma$ would only contain polynomials of (at most) a certain degree, and hence be a proper closed subspace of $C(I)$.

Fairly accessible classes of functions, such as polynomials and trigonometric polynomials (finite linear combinations of $\cos(m\cdot)$, $\sin(m\cdot)$), are also dense in $C(I)$ via the Stone-Weierstrass theorem. Recovering the right approximations would correspond to polynomial regression and Fourier decomposition tasks. What seems to make neural networks more attractive in practice is the *curse of dimensionality*; methods like polynomial regression are famously sensitive when n is large! A good choice of activation function σ can often make this regression task more stable.

References

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989.