UNIVERSAL APPROXIMATION

DEF (Universal approximator)

A class of functions $\mathcal{F}$ is a universal approximator over a compact set S if for every continuous function $h: S \mapsto \mathbb{R}$ and target accuracy $\varepsilon > 0$, $\exists f \in \mathcal{F}$ s.t.

$$\sup_{x \in S} |f(x) - h(x)| \leq \varepsilon.$$

Some tools from basic analysis: Uniform approximation in C^0

Let $C^0([a,b], \mathbb{R}) = \{h : [a,b] \to \mathbb{R} \;:\; h \text{ is continuous}\}$.

The first task is to approximate $h \in C^0$ by a smooth function f.

The ultimate smooth function is the polynomial $\Rightarrow$ polynomial approximation

Theorem A.1 (Weierstrass)

The set of polynomials is $\underline{\underline{\text{dense}}}$ in $C^0([a,b], \mathbb{R})$.

$\forall h \in C^0, \forall \varepsilon > 0, \; \exists$ polynomial $p(x)$ s.t.

$$\sup_{x \in [a,b]} |h(x) - p(x)| \leq \varepsilon.$$

Weierstrass is an approximation result merely on an interval. The next goal is to extend Weierstrass approx. theorem to functions defined on a metric space M.

Some definitions first.

DEF
- A subset $\mathcal{A}$ of $C^0(M, \mathbb{R}) =: C^0 M$ is a <u>function algebra</u> if it is closed under addition, scalar multiplication and function multiplication. $f, g \in \mathcal{A}, c \in \mathbb{R} \Rightarrow f+g, \; c \cdot f, \; f \cdot g \in \mathcal{A}$.
- A function algebra $\mathcal{A}$ vanishes at a point $x \in M$ if $f(x) = 0, \forall f \in \mathcal{A}$.
- A function algebra $\mathcal{A}$ separates points if $\forall x_1, x_2 \in M, x_1 \neq x_2,$ $\exists f \in \mathcal{A}$ s.t. $f(x_1) \neq f(x_2)$.

Some super easy examples.

(i) The set of polynomials is a function algebra.

The set of polynomials with a fixed degree is <u>not</u> a func. algebra.

(ii) $\{p : p \text{ is a polynomial with } p(0)=0\}$ vanishes at $x=0$.

(iii) the function algebra of all trigonometric polynomials separates points on $(0, 2\pi)$ and vanishes nowhere.

THEOREM A.2 (Stone–Weierstrass Theorem)

If M is a compact metric space and $\mathcal{A}$ is a function algebra in $C^0(M,\mathbb{R})$ that vanishes nowhere and separates points, then $\mathcal{A}$ is dense in $C^0 M$.

Universal approximation with neural networks

Let
$$\mathcal{F}_{\sigma,d,m} := \left\{ x \mapsto \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i) \; : \; a \in \mathbb{R}^m, \, W \in \mathbb{R}^{m\times d}, \, b \in \mathbb{R}^m \right\}$$

$$\mathcal{F}_{\sigma,d} := \bigcup_{m=1}^{\infty} \mathcal{F}_{\sigma,d,m} \longrightarrow \text{1-hidden-layer, unbounded width networks}$$

LEMMA 1 ($\mathcal{F}_{\cos,d}$) if $\sigma(z) = \cos(z)$, then $\mathcal{F}_{\sigma,d}$ is a function algebra.

Pf Consider
$$f(x) = \sum_{i=1}^{m} a_i \sigma(w_i^T x + b_i),$$
$$g(x) = \sum_{j=1}^{n} c_j \, \sigma(u_j^T x + v_j).$$

Then, apparently $f+g$, $\alpha \cdot f \in \mathcal{F}_{\cos,d}$. To check $f \cdot g \in \mathcal{F}_{\cos,d}$,

$$\begin{aligned}
f(x)g(x) &= \sum_{i,j} a_i c_j \cos(w_i^T x + b_i)\cos(u_j^T x + v_j) \\
&= \frac{1}{2}\sum_{i,j} a_i c_j \cos\left((w_i - u_j)^T x + b_i - v_j\right) \\
&\qquad - \frac{1}{2}\sum_{i,j} a_i c_j \, \sigma\left((w_i + u_j)^T x + b_i + v_j\right)
\end{aligned}$$

$\in \mathcal{F}_{\cos,d}$.

PROP 1 (Universality of $\mathcal{F}_{\cos,d}$)

$\mathcal{F}_{\cos,d}$ is universal over a compact set $M \subset \mathbb{R}^d$.

Pf: $\mathcal{F}_{\cos,d}$ is a function algebra in $C^0(M, \mathbb{R})$.

$\Rightarrow$ verify the conditions of Stone–Weierstrass Theorem.

(1) Vanishing nowhere: $\forall x \in M$, $\cos(0^T x) = \cos(0) = 1 \neq 0$

and apparently $\cos(0^T x) \in \mathcal{F}_{\cos,d}$ ✓

(2) $x \neq x'$, $f(z) = \cos\left(\dfrac{(z-x')^T(x-x')}{\|x-x'\|_2^2}\right) \in \mathcal{F}_{\cos,d}$.

Then, $f(x) = \cos(1)$, $f(x') = 1 \Rightarrow f(x) \neq f(x')$.

Thus, $\mathcal{F}_{\cos,d}$ separates points. $\blacksquare$

Remark: cos is quite an unusual activation function. We use it as an intermediate step.

The following result is the main result. It is a famous universal approximation theorem due to (Hornik et al., 1989).

THEOREM (Universal approximation, (Hornik, 1989))

Suppose $\sigma: \mathbb{R} \to \mathbb{R}$ is non-decreasing and $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to \infty} \sigma(z) = 1$.

Then, $\mathcal{F}_{\sigma,d}$ is universal.

Nomenclature:

$\sigma: \mathbb{R} \to \mathbb{R}$ continuous with $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to \infty} \sigma(z) = 1$ is called sigmoidal.

$\sigma: \mathbb{R} \to [0,1]$ s.t. σ is non-decreasing, $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to \infty} \sigma(z) = 1$ is called a squashing function. The original paper (Hornik et al., 1989) considers squashing activations, whereas we consider sigmoidals. Note that squashing functions have at most countably many discontinuities.

Proof of Theorem We will use two lemmas, which will be proved at the end.

Lemma 1 Let F be a non-decreasing sigmoidal, and σ be an arbitrary squashing function. Then, $\forall \varepsilon > 0$, $\exists H_\varepsilon \in \mathcal{F}_{\sigma,1}$ s.t.

$$\sup_{x \in \mathbb{R}} | F(x) - H_\varepsilon(x) | \leq \varepsilon.$$

Lemma 2 For every squashing function σ, $\forall \varepsilon > 0$, $\forall K > 0$, $\exists h \in \mathcal{F}_{\sigma,1}$ s.t. $\sup_{x \in [-K,K]} | h(x) - \cos(x) | < \varepsilon.$

Now, let's return to the proof of the theorem. By Prop 1, $\mathcal{F}_{\cos,d}$ is universal over a compact set $M \subset \mathbb{R}^d$. Thus, $\forall \varepsilon > 0$ and cont. h,

$$\sup_{x \in [0,1]^d} | h(x) - f(x) | < \varepsilon/2 \quad \text{for some } f \in \mathcal{F}_{\cos,d}.$$

Denote such $f \in \mathcal{F}_{\cos,d}$ by

$$f(x) = \sum_{i=1}^m a_i \cos(\omega_i^T x + b_i), \quad x \in [0,1]^d.$$

Set $K = \max_{1 \leq i \leq m} (\|\omega_i\|_2 + |b_i|)$, and $\varepsilon_0 = \dfrac{\varepsilon}{2 \cdot m \cdot \max_{1 \leq i \leq m} |a_i|}$. Then, by Lemma 2, for each $i \in [m]$, $\exists g_i \in \mathcal{F}_{\sigma,1}$ s.t.

$$\sup_{x \in [0,1]^d} | \cos(\omega_i^T x + b_i) - g_i(\omega_i^T x + b_i) | \leq \varepsilon_0.$$

Hence,

$$\sup_{x \in [0,1]^d} \left| h(x) - \sum_{i=1}^m a_i g_i(\omega_i^T x + b_i) \right| \leq \sup_{x \in [0,1]^d} | f(x) - h(x) | + \sup_{x \in [0,1]^d} \left| f(x) - \sum_{i=1}^m a_i g_i(\omega_i^T x + b_i) \right|$$

$$\leq \frac{\varepsilon}{2} + \sup_{x \in [0,1]^d} \left| f(x) - \sum_{i=1}^m a_i g_i(\omega_i^T x + b_i) \right|.$$

Now,

$$\left|f(x) - \sum_{i=1}^{m} a_i g_i(\omega_i^T x + b_i)\right| = \left|\sum_{i=1}^{m} a_i \left(\cos(\omega_i^T x + b_i) - g_i(\omega_i^T x + b_i)\right)\right|$$

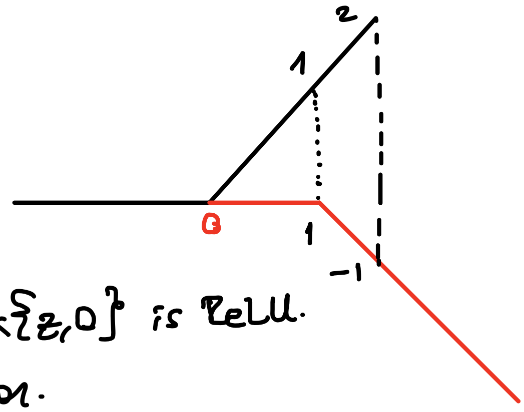
$$\leq \sum_{i=1}^{m} |a_i| . \left|\cos(\omega_i^T x + b_i) - g_i(\omega_i^T x + b_i)\right|$$

$$\leq \sum_{i=1}^{m} |a_i| . \varepsilon_0 = \sum_{i=1}^{m} |a_i| . \frac{\varepsilon}{2 . m . \|a\|_\infty}$$

$$\leq \frac{\varepsilon}{2} .$$

Hence,

$$\sup_{x \in [0,1]^d} \left|h(x) - \sum_{i=1}^{m} a_i g_i(\omega_i^T x + b_i)\right| \leq \varepsilon, \quad g_i \in \mathcal{F}_{\sigma,d}.$$



Important remarks

(1) ReLU is completely fine.

Let $\bar{\sigma}(z) = \sigma(z) - \sigma(z-1)$, $\sigma(z) = \max\{z, 0\}$ is ReLU.

Then, $\bar{\sigma}$ is a squashing function.

$z < 0 \Rightarrow \bar{\sigma}(z) = 0$

$z \in (0,1] \Rightarrow \bar{\sigma}(z) = z$

$z > 1 \Rightarrow \bar{\sigma}(z) = 1$.

$\sigma(z-1) = (z-1)\mathbb{1}\{z-1 \geq 0\}$

$\sigma(z) = z\,\mathbb{1}\{z \geq 0\}$

$$\sum_{i=1}^{m} a_i \bar{\sigma}(\omega_i^T x + b_i) = \sum_{i=1}^{m} a_i \sigma(\omega_i^T x + b_i) + \sum_{i=1}^{m} (-a_i)\sigma(\omega_i^T x + b_i - 1) \in \mathcal{F}_{\sigma,d,2m}$$

(2) Sigmoidal activation functions are universal also.

(3) One can use $\mathcal{F}_{\exp,d}$ as the intermediate function algebra to prove the result.

Exercise : Prove (2) and (3).

Supplementary Material

Proof of Lemma 1 :

Pick $\varepsilon \in (0,1)$. We must find (a_j, ω_j, b_j) for $j = 1, 2, \dots, m$ s.t.

$$\sup_{x \in \mathbb{R}} \left| F(x) - \sum_{i=1}^{m} a_i \, \sigma(\omega_j^T x + b_j) \right| \leq \varepsilon.$$

Here is the construction. Pick $m \in \mathbb{N}_+$ s.t. $\frac{1}{m+1} < \frac{\varepsilon}{2}$.

Pick $K \in \mathbb{R}_{>0}$ s.t. $\underbrace{\sigma(-M) < \frac{\varepsilon}{2(m+1)}, \quad \sigma(M) > 1 - \frac{\varepsilon}{2 \cdot (m+1)}}_{\text{such } M \text{ can be found by squashing properties of } \sigma.}$

For $i = 1, 2, \dots, m$, set $r_i = \sup \left\{ x \in \mathbb{R} : F(x) = \frac{i}{m+1} \right\}$, $a_i = \frac{1}{m+1}$,

$$r_{m+1} = \sup \left\{ x : F(x) = 1 - \frac{1}{2(m+1)} \right\}.$$

Such r_i, r_{m+1} exist since F is a nondecreasing sigmoidal.

For $p, q \in \mathbb{R}$, let $A_{p,q} : \mathbb{R} \to \mathbb{R}$ be the <u>unique</u> affine function s.t. $A_{p,q}(p) = K$ and $A_{p,q}(q) = -K$. (2 parameters in an affine function, 2 boundary conditions $\Rightarrow$ uniqueness).

Then, set

$$H_\varepsilon(x) = \sum_{i=1}^{m} a_i \cdot \sigma\left(A_{r_i, r_{i+1}}(x) \right).$$

It is straightforward to check $\sup_{x \in \mathbb{R}} |F(x) - H_\varepsilon(x)| < \varepsilon$. $\blacksquare$

Proof of Lemma 2 :

Let $F(z) \triangleq \frac{1 + \cos(z + \frac{3\pi}{2})}{2} \cdot \mathbb{1}\{-\pi/2 \leq z \leq \frac{\pi}{2}\} + \mathbb{1}\{z > \frac{\pi}{2}\}$.

This is a squashing function. (Gallant & White, 1988). Then, by adding and scaling a finite number of affinely shifted versions of the squashing function F, we can get $\cos(z)$ on any interval $[-K, K]$. Then, we obtain the result by a direct application of Lemma 1. $\blacksquare$