Approximation Theory

UNIVERSITY *of* WASHINGTON

Universal Approximation

Definition: A class of functions $\mathscr{F}$ is universal approximator over a compact set S (e.g., $[0,1]^d$), if for every continuous function g and a target accuracy $\epsilon > 0$, there exists $f \in \mathscr{F}$ such that

$$\sup_{x \in S} |f(x) - g(x)| \leq \epsilon$$

Stone-Weierstrass Theorem

Theorem: If $\mathscr{F}$ satisfies

1. Each $f \in \mathscr{F}$ is continuous.

2. $\forall x, \exists f \in \mathscr{F}, f(x) \neq 0$

3. $\forall x \neq x', \exists f \in \mathscr{F}, f(x) \neq f(x')$

4. $\mathscr{F}$ is closed under multiplication and vector space operations, $f_1, f_2 \in F, f_1 \cdot f_2 \in F$

Then $\mathscr{F}$ is a universal approximator:

$$\forall g : S \to R, \epsilon > 0, \exists f \in \mathscr{F}, \|f - g\|_\infty \leq \epsilon.$$

Example: cos activation

σ: activation

- $\mathcal{F}_{\sigma,d,m} = \{ x \mapsto \underbrace{a^T \sigma(Wx+b)}_{= \sum_{i=1}^{m} a_i \sigma(w_i^T x + b_i)}, a \in \mathbb{R}^m, W \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m \}$
- $\mathcal{F}_{\sigma,d} \triangleq \bigcup_{m \geq 0} \mathcal{F}_{\sigma,d,m}$

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

Pf: ① $\forall f \in \mathcal{F}$ is continuous

② $\forall x,\ \cos(0^T x) = 1 \neq 0$

③ $\forall x, x',$ choose some w s.t. $\cos(w^T x) \neq \cos(w^T x')$

④ $f, g \in \mathcal{F}_{\sigma,d} \Rightarrow f \cdot g \in \mathcal{F}_{\sigma,d}$

Recall $2\cos(y) \cdot \cos(z) = \cos(y+z) + \cos(y-z)$

$$2 \cdot \Big(\underbrace{\sum_{i=1}^{n} a_i \cos(u_i^T x + b_i)}_{f}\Big) \cdot \Big(\underbrace{\sum_{j=1}^{m} c_j \cos(v_j^T x + d_j)}_{g}\Big)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i c_j \left[\cos(u_i^T x + b_i + v_j^T x + d_j) + \cos(u_i^T x + b_i - v_j^T x - d_j)\right] \in \mathcal{F}_{\sigma,d}$$

Example: cos activation

Other Examples

Exponential activation

$a^T \exp(w^T x + b)$

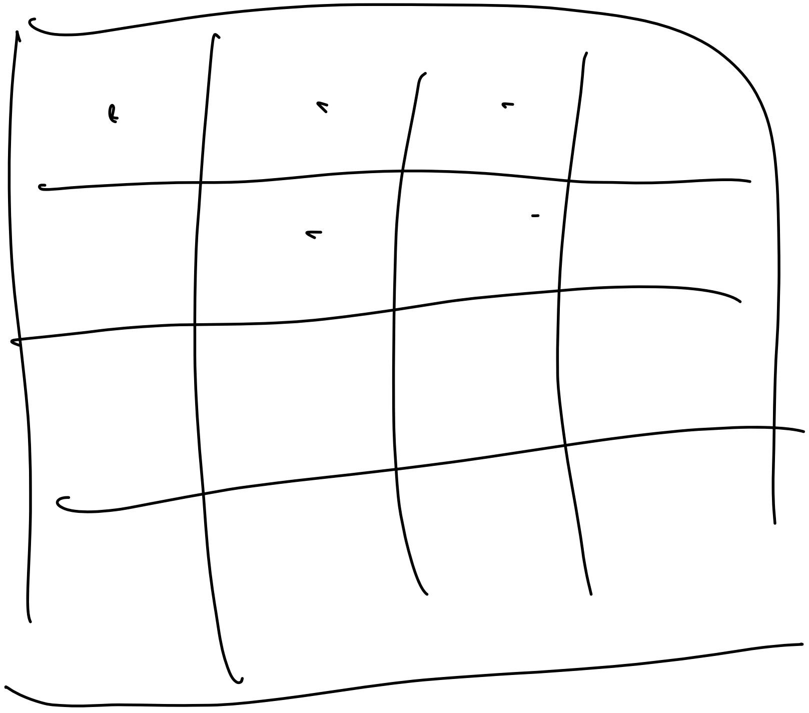
ReLU activation

Curse of Dimensionality

$O\left(\frac{1}{\delta^d}\right)$

$\Omega\left(\frac{1}{\delta^d}\right)$

- Unavoidable in the wors~~e~~t case



Barron's Theory

$v = \sum_{i=1}^{n} \lambda_i x_i$

λ_i: projection of v onto x_i

- Can we avoid the curse of dimensionality for "nice" functions?
- What are nice functions?
  - Fast decay of the Fourier coefficients

- Fourier basis functions:

$$\{e_w(x) = e^{i\langle w,x\rangle} = \underbrace{\cos(\langle w, x\rangle)}_{\text{Real}} + i \underbrace{\sin(\langle w, x\rangle)}_{\text{imag}} \mid w \in \mathbb{R}^d\}$$

- Fourier coefficient: $\hat{f}(w) = \int_{\mathbb{R}^d} f(x) e^{-i\langle w,x\rangle} dx$

projection

- Fourier integral / representation: $f(x) = \int_{\mathbb{R}^d} \hat{f}(w) e^{i\langle w,x\rangle} dw$

Barron's Theorem

Definition: The Barron constant of a function f is:

$$C \triangleq \int_{\mathbb{R}^d} \|w\|_2 |\hat{f}(w)| \, dw.$$

Theorem (Barron '93): For any $g : \mathbb{B}_1 \to \mathbb{R}$ where $\mathbb{B}_1 = \{x \in \mathbb{R} : \|x\|_2 \leq 1\}$ is the unit ball, there exists a 3-layer neural network f with $\boxed{O(\frac{C^2}{\epsilon})}$ neurons and sigmoid activation function such that

C is for g

$$\int_{\mathbb{B}_1} (f(x) - g(x))^2 dx \leq \epsilon.$$

Examples

- Gaussian function: $f(x) = (2\pi\sigma^2)^{d/2}\exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right)$

$$\hat{f}(w) = \exp\left(-2\pi\sigma^2\|w\|^2\right)$$

$$\text{let } Z = (2\pi\sigma^2)^{d/2}$$

$$C = \int_{\mathbb{R}^d} \|w\| \, |\hat{f}(w)| \, dw = Z \int \underset{\sim}{Z^{-1}} \|w\| \underline{\exp\left(-2\pi\sigma^2\|w\|^2\right)} dw$$

Gaussian distribution

$$Z \cdot \mathbb{E}\|X\|$$

$$\mathbb{E}\|X\| \le \sqrt{\mathbb{E}\|X\|^2}$$

$$\le Z \cdot \frac{1}{Z^{1/2}} \cdot \left(\frac{d}{4\pi^2\sigma^2}\right)^{\frac{1}{2}}$$

$$= Z^{\frac{1}{2}} \left(\frac{d}{4\pi^2\sigma^2}\right)^{\frac{1}{2}}$$

$$\text{if } 2\pi\sigma^2 < 1 \Rightarrow Z = O(1) \Rightarrow O(\sqrt{d})$$

$$\sigma \le \frac{1}{\sqrt{2\pi}}$$

- Other functions:
  - Polynomials
  - Function with bounded derivatives

Proof Ideas for Barron’s Theorem

Step 1: show any continuous function can be written as an infinite neural network with cosine-like activation functions.

(Tool: Fourier representation.)

Step 2: Show that a function with small Barron constant can be approximated by a convex combination of a small number of cosine-like activation functions.

(Tool: subsampling / probabilistic method.)

Step 3: Show that the cosine function can be approximated by sigmoid functions.

(Tool: classical approximation theory.)

Simple Infinite Neural Nets

Definition: An infinite-wide neural network is defined by a signed measure ν over neuron weights (w, b)

$$f(x) = \int_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sigma(w^\top x + b) d\nu(w, b).$$

$$\sum \underline{a_i} \sigma(u_i^T x + b_i)$$

Theorem: Suppose $g : \mathbb{R} \to \mathbb{R}$ is differentiable, if $x \in [0,1]$, then $g(x) = \int_0^1 \mathbf{1}\{x \geq b\} \cdot g'(b) db + g(0)$

Pf: by Fundamental Theorem of Calculu

$$g(x) = g(0) + \int_0^x g'(b) db$$

$$= g(0) + \int_0^1 \mathbb{1}\{x \geq b\} g'(b) db$$

Step 1: Infinite Neural Nets

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(w) \cdot e^{i\langle w,x\rangle} \, dw$$

The function can be written as

$$v(w)$$

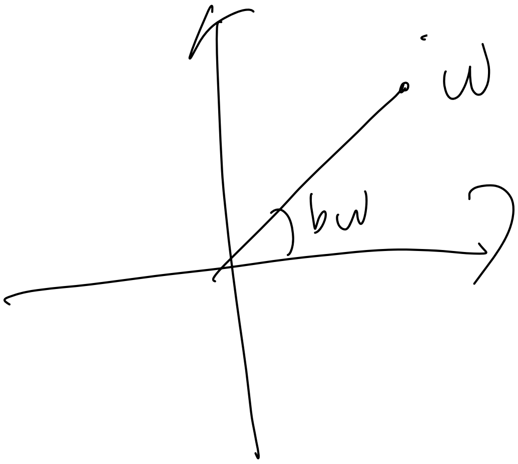
$$f(x) = f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| \underbrace{(\cos(b_w + \langle w, x\rangle) - \cos(b_w))}_{\sigma(wx+b\ )} dw.$$

pf: $f(x) = \int_{\mathbb{R}^d} \hat{f}(w) \cdot e^{i\langle w,x\rangle} dw$

$= \int_{\mathbb{R}^d} \hat{f}(w) dw + \int_{\mathbb{R}^d} \hat{f}(w)(e^{i\langle w,x\rangle} - 1) dw$

$(\hat{f}(w) = |\hat{f}(w)| e^{ib_w})$ $\quad = f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (e^{i(\langle w,x\rangle + b_w)} - e^{ib_w}) dw$

$(f \text{ is real})$ $\quad = f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (\cos(b_w + \langle w,x\rangle) - \cos(b_w)) dw$



$$w = |w| \cdot e^{i \cdot b_w}$$

Step 1: Infinite Neural Nets Proof

The function can be written as

$$f(x) = f(0) + \int_{\mathbb{R}^d} |\hat{f}(w)| (\cos(b_w + \langle w, x \rangle) - \cos(b_w)) dw.$$

Step 2: Subsampling

Writing the function as the expectation of a random variable:

$$f(x) = f(0) + \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|\|w\|_2}{\underline{C}} \left(\frac{C}{\|w\|_2} (\cos(b_w + \langle w, x \rangle) - \cos(b_w)) \right) dw.$$

We know $\displaystyle \int_{\mathbb{R}^d} \frac{|\hat{f}(w)| \, \|w\|_2 \, dw}{C} = 1$ ⇒ distribution over $w : D_w$

$$\mathbb{E}_{w \sim D_w}\left[\frac{C}{\|w\|_2} \left(\cos(b_w + \langle w, x \rangle) - \cos(b_w) \right) \right] = f(x) - f(0)$$

Step 2: Subsampling

Writing the function as the expectation of a random variable:

$$f(x) = f(0) + \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|\,\|w\|_2}{C} \left(\frac{C}{\|w\|_2} (\cos(b_w + \langle w, x \rangle) - \cos(b_w)) \right) dw.$$

Sample one $w \in \mathbb{R}^d$ with probability $\frac{|\hat{f}(w)|\,\|w\|_2}{C}$ for r times.

draw $\{w_1, \ldots, w_r\}$

if r is large enough: $O\left(\frac{C^2}{\varepsilon}\right)$

$$\Rightarrow \frac{1}{r} \sum_{j=1}^{r} \frac{C}{\|w_j\|} \cos(b_{w_j} + \langle w_j, x \rangle) - \cos(b_{w_j}) \approx f(x) - f(0$$

Step 3: Approximating the Cosines

Lemma: Given $g_w(x) = \frac{C}{\|w\|_2}(\cos(b_w + \langle w, x \rangle) - \cos(b_w))$, there exists a 2-layer neural network f_0 of size $O(1/\epsilon)$ with sigmoid activations, such that $\sup_{x \in [-1,1]} |f_0(y) - h_w(y)| \leq \epsilon$.

① use threshold function → cos

② sigmoid → threshold

Depth Separation

So far we only talk about 2-layer or 3-layer neural networks.

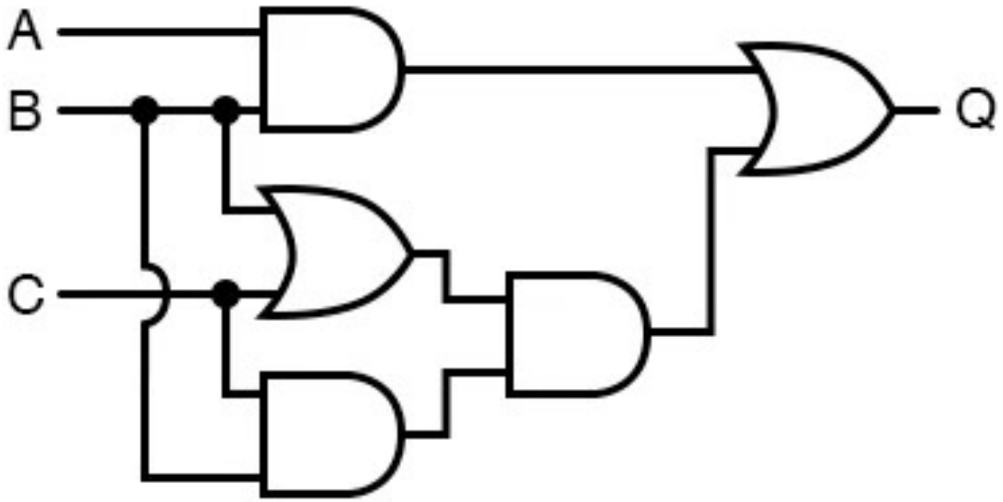
Why we need **Deep** learning?

Can we show deep neural networks are **strictly** better than shallow neural networks?

A brief history of depth separation

Early results from theoretical computer science

Boolean circuits: a directed acyclic graph model for computation over binary inputs; each node (“gate”) performs an operation (e.g. OR, AND, NOT) on the inputs from its predecessors.



A brief history of depth separation

Early results from theoretical computer science

Boolean circuits: a directed acyclic graph model for computation over binary inputs; each node (“gate”) performs an operation (e.g. OR, AND, NOT) on the inputs from its predecessors.

Depth separation: the difference of the computation power: shallow vs deep Boolean circuits.

Håstad (’86): parity function cannot be approximated by a small constant-depth circuit with OR and AND gates.

Modern depth-separation in neural networks

- **Related architectures / models of computation**
  - Sum-product networks [Bengio, Delalleau ’11]

- **Heuristic measures of complexity**
  - Bound of number of linear regions for ReLU networks [Montufar, Pascanu, Cho, Bengio ‘14]

- **Approximation error**
  - A small deep network cannot be approximated by a small shallow network [Telgarsky ’15]

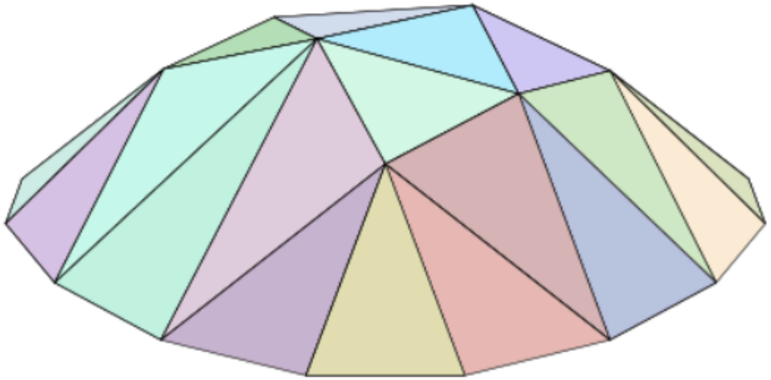
Shallow Nets Cannot Approximate Deep Nets

Theorem (Telgarsky ’15): For every $L \in \mathbb{N}$, there exists a function $f : [0,1] \to [0,1]$ representable as a network of depth $O(L^2)$, with $O(L^2)$ nodes, and ReLU activation such that, for every network $g : [0,1] \to \mathbb{R}$ of depth L and $\leq 2^L$ nodes, and ReLU activation, we have

$$\int_{[0,1]} |f(x) - g(x)|\, dx \geq \frac{1}{32}.$$

Intuition

A ReLU network f is piecewise linear, we can subdivide domain into a finite number of polyhedral pieces $(P_1, P_2, \ldots, P_N)$ such that in each piece, f is linear: $\forall x \in P_i, f(x) = A_i x + b_i$.



Deeper neural networks can make exponentially more regions than shallow neural networks.

Make each region has different values, so shallow neural networks cannot approximate.

Benefits of depth for smooth functions

Theorem (Yarotsky '15): Suppose $f : [0,1]^d \to \mathbb{R}$ has all partial derivatives of order r with coordinate-wise bound in $[-1,1]$, and let $\epsilon > 0$ be given. Then there exists a $O(\ln \frac{1}{\epsilon})$ - depth and $\left(\frac{1}{\epsilon}\right)^{O(\frac{d}{r})}$ -size network so that $\sup_{x \in [0,1]^d} |f(x) - g(x)| \leq \epsilon$.

Remarks

- All results discussed are existential: they prove that a good approximator exists. Finding one efficiently (e.g., using gradient descent) is the next topic (optimization).
- The choices of non-linearity are usually very flexible: most results we saw can be re-proven using different non-linearities.
- There are other approximation error results: e.g., deep and narrow networks are universal approximators.
- Depth separation for optimization and generalization is widely open.

Recent Advances in Representation Power

- Analyses of different architectures
  - Graph neural network
  - Attention-based neural network
- Separation between transformers and RNNs
- Finite data approximation
- In-context learning for specific tasks
- Chain-of-thought
- …