

Proposal Due 4/8 11:59PM
on Canvas
1-1.5 page

Approximation Theory

Specific Setups

- “Average” approximation: given a distribution μ

$$\|f - g\|_{\mu} = \int_x |f(x) - g(x)| d\mu(x)$$
$$\|f - g\|_{\mu}^2 = \int_x |f(x) - g(x)|^2 d\mu(x)$$

- “Everywhere” approximation

$$\|f - g\|_{\infty} = \sup_x |f(x) - g(x)| \geq \|f - g\|_{\mu}$$

Multivariate Approximation

Theorem: Let g be a continuous function that satisfies $\|x - x'\|_\infty \leq \delta \Rightarrow |g(x) - g(x')| \leq \epsilon$ (Lipschitzness). Then there exists a **3-layer ReLU neural network** with

$O(\frac{1}{\delta^d})$ nodes that satisfy

$$\int_{[0,1]^d} |f(x) - g(x)| dx = \|f - g\|_1 \leq \epsilon$$

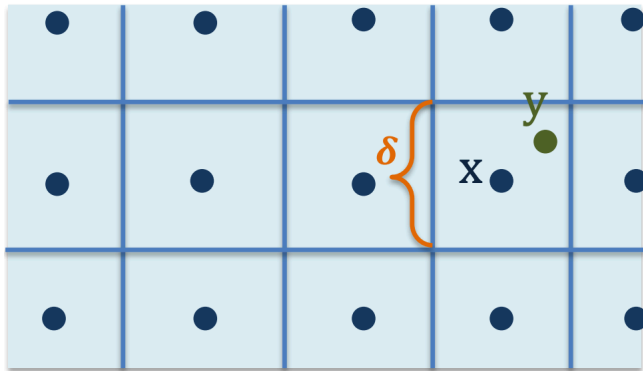
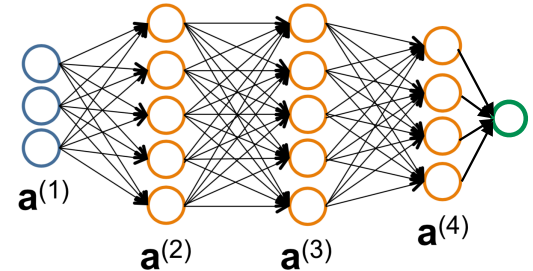


Figure credit to Andrej Risteski



Universal Approximation

Definition: A class of functions \mathcal{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 $f \in \mathcal{F}$ such that

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

Stone-Weierstrass Theorem

Theorem: If \mathcal{F} satisfies

1. Each $f \in \mathcal{F}$ is continuous.
2. $\forall x, \exists f \in \mathcal{F}, f(x) \neq 0$
3. $\forall \underline{x \neq x'}, \exists f \in \mathcal{F}, \underline{f(x) \neq f(x')}$
4. \mathcal{F} is closed under multiplication and vector space operations, $f_1, f_2 \in \mathcal{F} \implies f_1 + f_2 \in \mathcal{F}$

Then \mathcal{F} is a universal approximator:

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

Example: cos activation

$$- \mathcal{F}_{\cos, d, m} = \left\{ x \mapsto a^T \cos(Wx + b), \begin{matrix} a \in \mathbb{R}^m \\ W \in \mathbb{R}^{m \times d} \\ b \in \mathbb{R}^m \end{matrix} \right\}$$

input dim \nwarrow \swarrow # of nodes

$$- \mathcal{F}_{\cos, d} \triangleq \bigcup_{m \geq 1} \mathcal{F}_{\cos, d, m}$$

$\mathcal{F}_{\cos, d}$ is universal approximator

- Pf:
- (1) $f \in \mathcal{F}_{\cos, d}$ is continuous
 - (2) $\forall x, \exists f$ s.t. $f(x) \neq 0$, ($a=1, W=0, b=0$)
 - (3) $f, g \in \mathcal{F}_{\cos, d} \Rightarrow fg \in \mathcal{F}_{\cos, d}$

$$f \in \mathcal{F}_{\cos, d, m}$$

$$g \in \mathcal{F}_{\cos, d, m}$$

$$\begin{aligned}
 & \cdot 2 \cos y \cos z = \cos(y+z) + \cos(y-z) \\
 & \left[\sum_{i=1}^m a_i \cos(w_i^T x + b_i) \right] \cdot \left[\sum_{j=1}^m c_j \cos(v_j^T x + d_j) \right] \\
 & = \sum_{i=1}^m \sum_{j=1}^m \frac{a_i c_j}{2} \left[\cos((w_i + v_j)^T x + b_i + d_j) + \cos((w_i - v_j)^T x + b_i - d_j) \right] \\
 & \in \mathcal{F}
 \end{aligned}$$

Example: cos activation

(4) $\forall x \neq x', \exists f(x) \neq f(x')$

define $f(z) = \cos\left(\frac{(z-x')^T(x-x')}{\|x-x'\|_2^2}\right) \in \mathcal{F}$

$$f(x) = \cos(1) \neq$$

$$f(x') = \cos(0)$$

□

Other Examples

Exponential activation

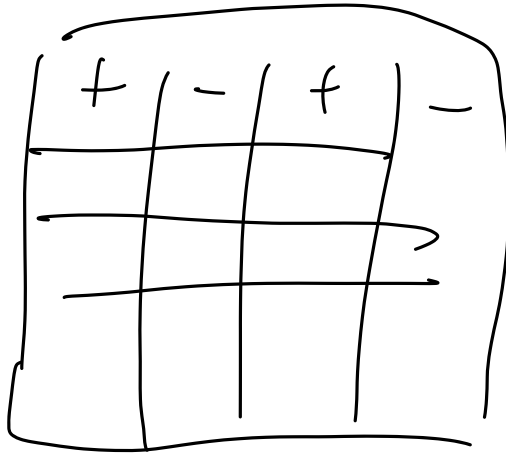
$$F_{\exp, d}$$

ReLU activation

Then σ is continuous, $\lim_{z \rightarrow -\infty} \sigma(z) = 0$
 $\lim_{z \rightarrow \infty} \sigma(z) = 1$ $F_{\sigma, d}$ is universal
Use ReLU to approximate
 $\hat{\sigma} \leftarrow \hat{\sigma}(z) - \hat{\sigma}(z-1)$

Curse of Dimensionality

- Unavoidable in the worse case



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

non-parametric

Barron's Theory

time \longleftrightarrow frequency

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

- Fourier basis functions:

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

a set of functions

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

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

χ $\begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$ $\chi = \sum \lambda_i V_i$
 $\{V_i\}$ is a basis

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 $g : \mathbb{B}_1 \rightarrow \mathbb{R}$ 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 $O(\frac{C^2}{\epsilon})$ neurons and (small

sigmoid activation function such that

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

Examples

Need $O\left(\frac{d}{\epsilon}\right)$

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

$$\hat{f}(x) = \exp(-2\pi\sigma^2 \|u\|_2^2)$$

let $Z = (2\pi\sigma^2)^{-d/2}$ normalizing constant

$$C = \int \|u\| | \hat{f}(u) | du = Z \int \underbrace{Z^{-1} | \hat{f}(u) | \|u\|}_{\text{red underline}} du$$

$$\left(\mathbb{E}[\|u\|] \leq \sqrt{\mathbb{E}[\|u\|^2]} \right) \leq Z \left(\int \underbrace{Z^{-1} | \hat{f}(u) |}_{\text{red underline}} \frac{\mathbb{E}[\|u\|]}{\|u\|^2} du \right)^{1/2}$$

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

57 $2\pi\sigma^2 \geq 1$

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.

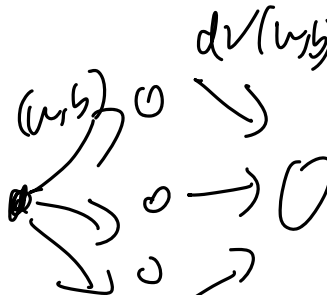
(Tool: classical approximation theory.)

Simple Infinite Neural Nets

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

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

first layer



Theorem: Suppose $g : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable, if

$$x \in [0, 1], \text{ then } g(x) = \int_0^1 \underbrace{\mathbf{1}\{x \geq b\}}_b \cdot \underbrace{g'(b) db}_{d\nu(b)} + g(0)$$

Pf: by Fundamental Theorem of calculus

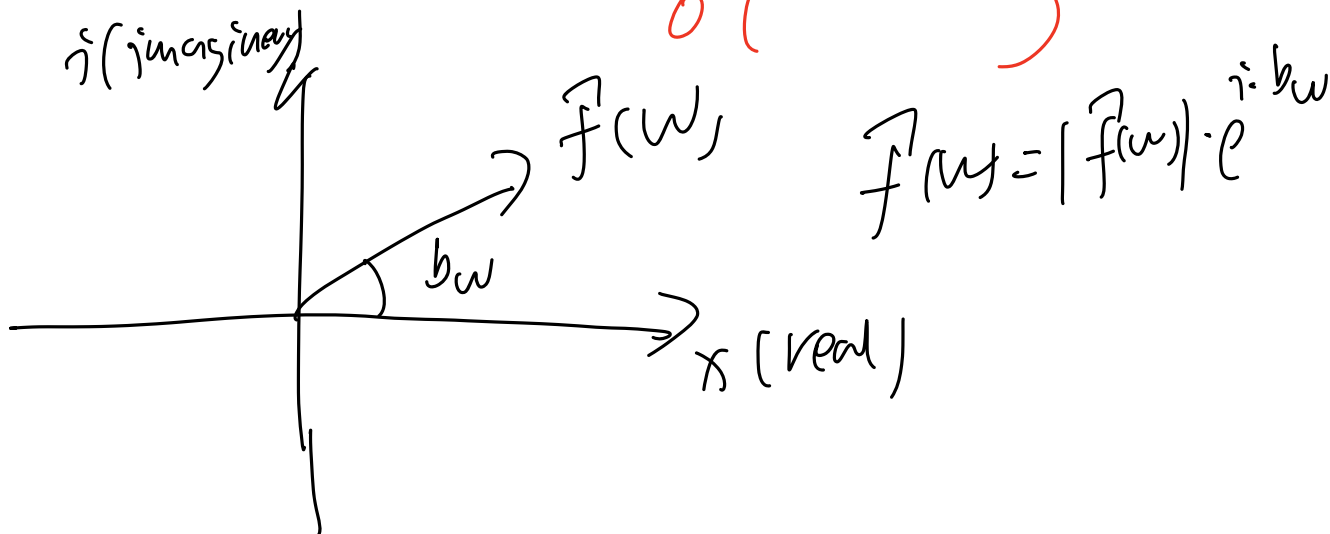
$$\begin{aligned} g(x) &= g(0) + \int_0^x g'(b) db \\ &= g(0) + \int_0^1 \mathbf{1}\{x \geq b\} g'(b) db \end{aligned}$$

Step 1: Infinite Neural Nets

$$d\nu(w)$$

The function can be written as

$$f(x) = f(0) + \int_{\mathbb{R}^d} \underbrace{|\hat{f}(w)|}_{\substack{\uparrow \\ \text{imaginary}}} \underbrace{(\cos(b_w + \langle w, x \rangle) - \cos(b_w))}_{\substack{\uparrow \\ \text{real}}} dw.$$



Step 1: Infinite Neural Nets Proof

The function can be written as

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

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

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

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

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

Step 2: Subsampling

Writing the function as the expectation of a random variable: (*)

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

$\int_{\mathbb{R}^d} \hat{f}(w) dw = C$ distribution over w D_w

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

Step 2: Subsampling

Writing the function as the expectation of a random variable:

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

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

$$\{w_1, \dots, w_r\}$$

$$f(0) + \frac{1}{r} \sum_{i=1}^r \frac{C}{\|w_i\|_2} (\cos(b_{w_i} + \langle w_i, x \rangle) - \cos(b_{w_i}))$$

$$\approx f(x)$$

by concentration

$$r \approx O\left(\frac{C^2}{\epsilon^2}\right)$$

Step 3: Approximating the Cosines

id function

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 $\sup_{x \in [-1, 1]} |f_0(y) - h_w(y)| \leq \epsilon$.

$$(1) \quad g_w(x) = \frac{C}{\|w\|_2} (\cos(b_w + \|w\| \underbrace{\langle \frac{w}{\|w\|}, x \rangle}_y) - \cos(b_w))$$

$\stackrel{D}{=} h_w(y)$

- (2) use id approximation with threshold
- (3) use sigmoid to approximate threshold

Depth Separation

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

Why we need **Deep** learning?

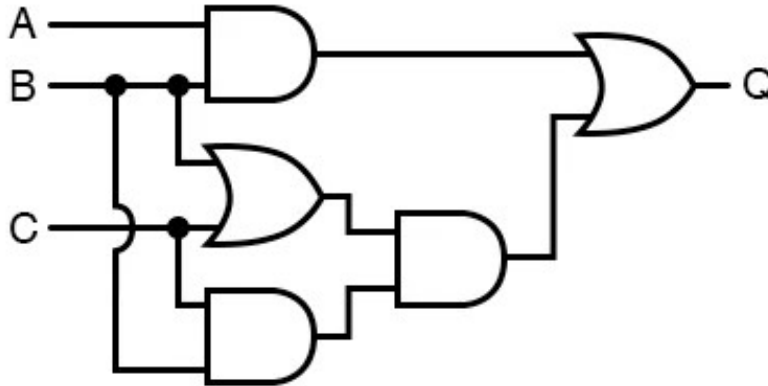
Can we show deep neural networks are **strictly** better than shallow neural networks?

← approximately
opt
generalization

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]

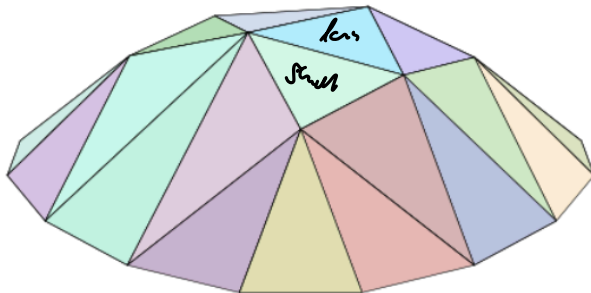
Shallow Nets Cannot Approximate Deep Nets

Theorem (Telgarsky '15): For every $L \in \mathbb{N}$, **there exists** a function $f : [0,1] \rightarrow [0,1]$ representable as a network of **depth $O(L^2)$, with $O(L^2)$ nodes**, and ReLU activation such that, for **every network $g : [0,1] \rightarrow \mathbb{R}$** 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, \dots, 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 $f : [0,1]^d \rightarrow \mathbb{R}$ 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 $\sup_{x \in [0,1]^d} |f(x) - g(x)| \leq \epsilon$. $d \leq \nu$

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.