Lecture 3, Monday, April 8, 2024

Approximation Theory of Neural Networks

- Introduction
  - An example
  - Review of two classical approximation Thms (Stone-Weierstrass, Kolmogrov Superposition Thm)
- Universal Approximation Thms.
- Error bounds of NN approximations
- Other topics

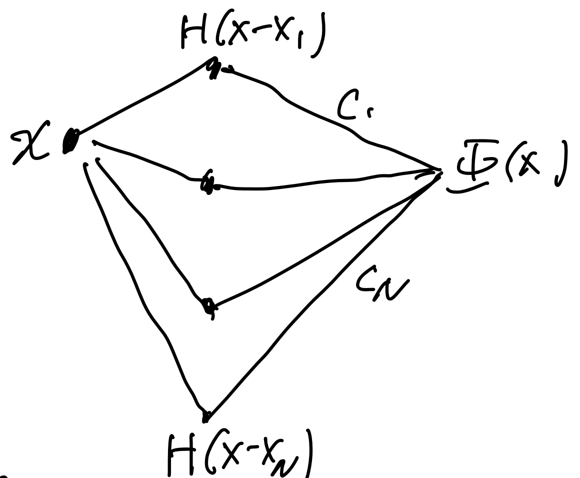
<u>An example</u> Given $g \in C([0,1])$. Construct a simple NN, $\Phi \in NN(H, L=2, N_0=1, N_L=1)$, to approximate g. Here,

$$H(x) = \begin{cases} 0 & \text{if } x < 0. \\ 1 & \text{if } x \geq 0. \end{cases}$$ — the Heaviside function. ($H(0)=1$)

<u>Thm</u> $\forall \varepsilon > 0$, $\exists N \in \mathbb{N}$ ($N >> 1$), $\exists\, 0 \leq x_1 < \cdots < x_N \leq 1$, $\exists\, c_1, \ldots, c_N \in \mathbb{R}$, s.t. $|\Phi_N(x) - g(x)| < \varepsilon \quad \forall x \in [0,1]$,

where

$$\Phi(x) = \Phi_N(x) = \sum_{j=1}^{N} c_j \cdot H(x - x_j) \in NN(H, 2, 1, 1), \quad x \in [0,1].$$



$-x_j = \theta_j$: threshold / bias.

c_j — weights in the output

$$A_1 = {}_N\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_1, \quad b_1 = \begin{bmatrix} -x_1 \\ \vdots \\ -x_N \end{bmatrix}.$$

$$A_2 = {}_1[c_1 \cdots c_N]_N, \quad b_2 = 0 \in \mathbb{R}.$$

<u>Proof</u> Since $g \in C([0,1])$ is unif. cont. on $[0,1]$, $\exists\, \delta > 0$ s.t. $|x-y| < \delta \Rightarrow |g(x) - g(y)| < \varepsilon$. Choose $N \in \mathbb{N}$ s.t. $N > \frac{1}{\delta}$.

Define $x_j = (j-1)/N$, $j = 1, \cdots, N$. On each $[x_j, x_{j+1}]$,

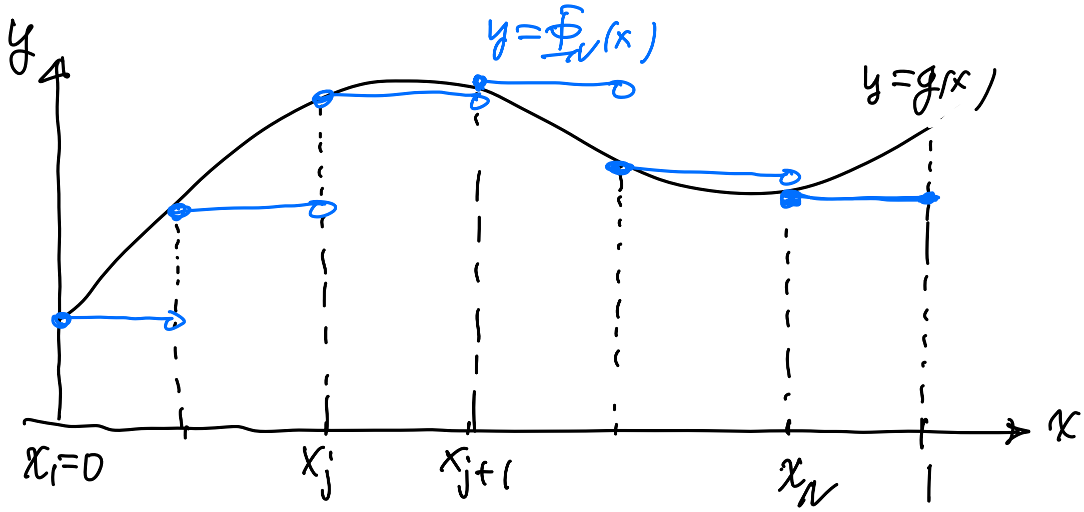
$$|g(x) - g(x_j)| < \varepsilon \qquad \forall x \in [x_j, x_{j+1}].$$

Define $c_1 = g(x_1)$, $c_j = g(x_j) - g(x_{j-1})$, $j = 2, \cdots, N$. Then,

$$\Phi_N(x) = \sum_{j=1}^{N} c_j H(x - x_j)$$

$$= \begin{cases} g(x_1) & \text{if } x \in [x_1, x_2), \\ g(x_2) & \text{if } x \in [x_2, x_3), \\ \cdots\cdots & \\ g(x_{j+1}) & \text{if } x \in [x_j, x_{j+1}], \\ \cdots\cdots & \\ g(x_N) & \text{if } x \in [x_N, 1]. \end{cases}$$

Then $|\Phi_N(x) - g(x)| < \varepsilon$. $\forall x \in [0, 1]$. QED



The First Weierstrass Approximation Theorem. Let $a, b \in \mathbb{R}$ with $a < b$. Let $f \in C([a,b])$ and $\varepsilon > 0$. Then there exists a polynomial p such that

$$\max_{a \le x \le b} |f(x) - p(x)| < \varepsilon. \qquad (*)$$

Remarks.

① Define for $g \in C([a,b])$,

$$\|g\| = \|g\|_\infty = \|g\|_{C([a,b])} = \|g\|_C = \max_{a \le x \le b} |g(x)|.$$

Then $(C([a,b]), \|\cdot\|)$ is a Banach space. The inequality $(*)$ is the same as $\|f - p\| < \varepsilon$.

- Equivalently $\lim_{n\to\infty} E_n(f)=0$, where
  $E_n(f)=\min_{q\in\mathcal{P}_n}\|f-q\|$, (min is attained uniquely).
  $\mathcal{P}_n=\{$all polynomials of $\deg\le n\}$.
- The theorem provides no info about p.

<u>The Second Weierstrass Approximation Theorem</u> Let $f\in C_{2\pi}$ and $\varepsilon>0$ then there exists a trigonometric polynomial T such that
$$\max_{-\pi\le x\le\pi}|f(x)-T(x)|<\varepsilon.$$

<u>The Stone-Weierstrass Theorem</u> Let X be a compact Hausdorff space and $C(X,\mathbb{R})$ or $C(X)$ the space of all real continuous functions equipped with the uniform norm. Let $\mathcal{A}\subseteq C(X,\mathbb{R})$. Assume
- $\mathcal{A}$ is a subalgebra;
- $\mathcal{A}$ separates points of X; and
- $\mathcal{A}$ contains the constant functions.

Then $\overline{\mathcal{A}}=C(X,\mathbb{R})$ (i.e., $\mathcal{A}$ is dense in $C(X,\mathbb{R})$).

<u>Remarks</u>
- "$\mathcal{A}$ is a subalgebra" means: $\mathcal{A}$ is a vector subspace of $C(X,\mathbb{R})$ and $\mathcal{A}$ is closed in multiplication: $f,g\in\mathcal{A}\Rightarrow f+g\in\mathcal{A}$; $f\in\mathcal{A}$, $\alpha\in\mathbb{R}$ $\Rightarrow \alpha f\in\mathcal{A}$; $f,g\in\mathcal{A}\Rightarrow f\cdot g\in\mathcal{A}$.
- $\mathcal{A}$ separates points of X: $\forall x,y\in X$, $x\ne y$. $\exists f\in\mathcal{A}$ s.t. $f(x)\ne f(y)$.
- $\overline{\mathcal{A}}$ is the closure of $\mathcal{A}$ w.r.t. the uniform

norm. $\overline{\mathcal{A}} = C(X, \mathbb{R})$ means that $\forall f \in C(X, \mathbb{R})$ there exist $g_k \in \mathcal{A}$ $(k = 1, 2, \cdots)$ such that $g_k \to f$ as $k \to \infty$. i.e.,

$$\|f - g_k\| = \max_{x \in X} |f(x) - g_k(x)| \to 0 \quad \text{as } k \to \infty.$$

(·) You may consider X to be a compact subset of $\mathbb{R}^n$. Compact here = closed + bounded.

<u>Example</u> $X = \prod_{j=1}^{n} [a_j, b_j]$. $(a_j, b_j \in \mathbb{R},\ a_j < b_j;\ j = 1, \cdots, n)$

or $X = \overline{co(X_1, \cdots, X_m)}$ the convex hull of m points in $\mathbb{R}^n$ $(X_1, \cdots, X_m \in \mathbb{R}^n)$. $\mathcal{A} = \{$all real multivariable polynomials $p = p(x_1, \cdots, x_n)\}$.

Then $\overline{\mathcal{A}} = C(X, \mathbb{R})$, i.e., $\forall f \in C(X)$ $\forall \varepsilon > 0$, there exists a polynomial $p: \mathbb{R}^n \to \mathbb{R}$ s.t.

$$|f(x) - p(x)| < \varepsilon \quad \forall x \in X.$$

The classical Weierstrass Thm.

<u>Corollary</u>. Let $X = \prod_{j=1}^{n} [a_j, b_j]$ with all $a_j, b_j \in \mathbb{R}$ and $a_j < b_j$ $(j = 1, \cdots, n)$. Let $f \in C(X)$ and $\varepsilon > 0$. Then, $\exists N \in \mathbb{N}$, and $f_{ij} \in C([a_j, b_j])$ $(i = 1, \cdots, N;\ j = 1, \cdots, n)$ such that

$$\max_{x = (x_1, \cdots, x_n) \in X} \left| f(x) - \sum_{i=1}^{N} \prod_{j=1}^{n} f_{ij}(x_j) \right| < \varepsilon.$$

<u>Proof</u> Let $\mathcal{A} = \left\{ \sum_{i=1}^{N} \prod_{j=1}^{n} g_{ij}(x_j) : g_{ij} \in C([a_j, b_j]),\ j = 1, \cdots, n,\ i = 1, \cdots, N,\ N = 1, 2, \cdots \right\}$.

Apply the Stone-Weierstrass Thm. <u>QED</u>

(Fill in details!)

Remark. We can choose $f_{ij} \in \mathcal{P}$ (the set of all one-variable real polynomials).

The Kolmogorov Thm (or Kolmogorov–Arnold Thm)

Kolmogorov, 1956, Arnold 1957, Kolmogorov 1957.

Thm (Kolmogorov 1957) For any integer $n \ge 2$ and and $f \in C([0,1]^n)$ there exist $\psi^{p,q} \in C([0,1])$ ($p = 1, \cdots, n$, $q = 1, \cdots, 2n+1$) and $\chi_q \in C(\mathbb{R})$ ($q = 1, \cdots, 2n+1$) such that

$$f(x_1, \cdots, x_n) = \sum_{q=1}^{2n+1} \chi_q \left(\sum_{p=1}^{n} \psi^{pq}(x_p) \right) \quad \forall x = (x_1, \cdots, x_n) \in [0,1]^n.$$

Remarks

- Originated from Hilbert's 13th prob. (1900).
- Constructive proof is given by Braun and Griebel (Constr. Approx. 2009), after Sprecher (1996, 1997) and Köppen (2002).

Universal Approximations

Question: Given a function/map: $f: X \to \mathbb{R}^m$, are there NNWs Φ_k ($k = 1, 2, \cdots$) such that $\Phi_k \to f$? Here $X \subseteq \mathbb{R}^d$ and the convergence is w.r.t. some norm or metric or some topology.

More questions

- What is the class of functions that can be approximated well by NNWs?
- The norm/metric? Uniform norm?

$L^1, L^2, L^p\ (1\le p<\infty), L^\infty$? $W^{k,p}$?

(i) Sequence of NNs w.r.t. depth $L>>1$, ? with $W>>1$?

(i) NN approximations vs. classical approximations (by polynomials, piecewise polynomials, trigonometric polynomials, splines, wavelets, finite elements, etc. and data fitting, e.g., interpolation, least-squares, etc.) ?

<u>Some observations / considerations</u>

(i) Only need to consider the output dimension $= 1$? Yes, for fixed L. But generally?

Suppose $\Phi^{(j)} \in NN(\sigma, L, N_0, N_L = 1)$ and $\Phi^{(j)} \approx f^{(j)}: X (\subseteq \mathbb{R}^{N_0}) \to \mathbb{R}^1$, $j=1,\cdots,m$. Then by parallelization with shared inputs, we can construct $\Phi \in NN(\sigma, L, N_0, N_L = m)$ such that $(\Phi(x))_j = \Phi^{(j)}(x)\ \forall x \in X$ for $j=1,\cdots,m$. Thus, $\Phi \approx f$ with $f = \begin{bmatrix} f^{(1)} \\ \vdots \\ f^{(m)} \end{bmatrix}$.

In deep NN approximations, we may have $\Phi^{(j)} \in NN(\sigma, L_j, N_0, N_{L_j} = 1)$ $(j=1,\cdots,m)$. <u>How to construct Φ from these $\Phi^{(j)}$ so that Φ can approximate $f: X (\subseteq \mathbb{R}^{N_0}) \to \mathbb{R}^{\max(N_{L_1},\cdots,N_{L_m})}$</u>? This leads to the question: <u>how to add one layer to an existing NN to get a new NN which is the same function?</u> For same σ, it may be fine. But in general, unknown or may not be possible.

The worst is: for each component of $f: X \to \mathbb{R}^m$, we approximate it by a sequence of NNs.

⊙ Any polynomial will not work (as an activation function). Since, if σ is a polynomial of degree n, then any $\Phi \in NN(\sigma, N_0=1, N_L=1)$ is also a polynomial of degree $\le n$. Therefore $NN(\sigma, N_0=1, N_L=1)$ is a finite-dimensional space and cannot approxim $C([0,1])$ which is infinitely dimensional.

⊙ Proof of universal approximations — abstract vs. constructive.

<u>Early works</u>

⊙ G. Cybenko, Math. Control Signals Systems, 1989.

⊙ K.-I. Funahashi, Neural Networks, 1989.

⊙ K. Hornik, M. Stinchcombe, and H. White, Neural Networks, 1989.

⊙ K. Hornik, Neural Networks, 1991.

⊙ Leshno et al. Neural Networks. 1993

⊙ Jones 1990 constructive proof.

⊙ Carroll Dickson (1990) Also constructions.

<u>Some notations / definitions</u>.

⊙ <u>The space $C(K)$</u>.

⊙ $d \in \mathbb{N}$: input dimension.

⊙ $K \subseteq \mathbb{R}^d$: compact (i.e., bounded + closed, e.g., $K = [0,1]^d$, $K =$ a closed ball in $\mathbb{R}^d$, etc.)

$C(K) = \{$ all real-valued continuous functions $K \to \mathbb{R}\}$.

$\|f\| = \|f\|_C = \|f\|_{C(K)} = \|f\|_\infty = \max_{x \in K} |f(x)|$.

Unif. norm. or max. norm.

Proposition $(C(K), \|\cdot\|)$ is a Banach space.

Let $\mathcal{A} \subseteq C(K)$. $\mathcal{A}$ is dense in $C(K)$ if: $\forall f \in C(K)$ there exist $\Phi_k \in \mathcal{A}$ $(k = 1, 2, \cdots)$ such that $\|\Phi_k - f\| \to 0$ as $k \to \infty$. Equivalently, $\forall f \in C(K)$ $\forall \varepsilon > 0$ $\exists \Phi \in \mathcal{A}$ such that $\|\Phi - f\| < \varepsilon$. Equivalently, $f \in \overline{\mathcal{A}}$, the closure of $\mathcal{A}$ in $C(K)$.

Remark A NN function is defined on the entire space of the input space $\mathbb{R}^d$. If we want to approximate $f \in C(K)$, $K \subseteq \mathbb{R}^d$: compact, we can approximate f by some $\tilde{f} \in C(Q)$, $Q = [a, b]^d \supseteq K$, $-\infty < a < b < \infty$, by using mollifiers and approximation to get $\hat{f} \in C_c(\mathbb{R}^d)$, even that $\hat{f} \in C_c^\infty(\mathbb{R}^d)$, $\operatorname{supp} \hat{f} \subseteq Q$. So, we can consider Q instead of K.

Question: If NNs approx. $C(K)$, can they approx $C(K')$ for any different K'? (K, K': compact).

$NN(\sigma, L, d, 1) =$ all NNs with the activation function σ, depth L (# hidden layers $= L-1$), input dimension d, and output dimension 1.

A Universal Approximation Theorem

Theorem (Cybenko 1989, Funahashi 1989, and Hornik, Stinchcombe, and White 1989) Let $\sigma \in C(\mathbb{R})$ be such that $\sigma(-\infty) = 0$ and $\sigma(+\infty) = 1$. Then $NN(\sigma, 2, d, 1)$ is dense in $C(K)$ for any compact $K \subseteq \mathbb{R}^d$ and any $d \in \mathbb{N}$.

Remarks

⊙ Three papers published in the same year! Proved the same result using different methods — so there are three different methods. Note: Funahashi (1989) and Hornik, Stinchcombe, & White (1989) require σ to be increasing. Also, Carroll and Dickinson (1990) gave a constructive proof of the result using Radon transforms.

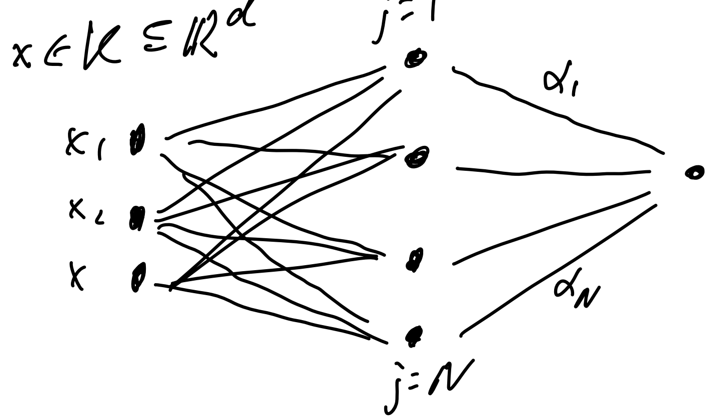
Here, we detail Cybenko's proof. We also sketch Funahashi's proof which is based on an integral formula of Irie-Miyake (1988). We also briefly describe the ideas of proof by Hornik, Stinchcombe, and White.

⊙ Note that the ReLU is not covered in this theorem. Hornik (1991) extended the result to any $\sigma \in C(\mathbb{R})$ that is bounded and non constant. The most general result is given by Leshno, Lin, Pinkus, and Schocken (1993): $\sigma \in C(\mathbb{R})$ leads to the universal approximation property $\Longleftrightarrow$ σ is not a polynomial. We will sketch the proof of this general result.

⊙ Representation of $G \in NN(\sigma, 2, d, 1)$:

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + \theta_j) + \beta, \quad \forall x \in \mathbb{R}^d, \quad (*)$$

where $\alpha_j \in \mathbb{R}$, $w_j \in \mathbb{R}^d$, $\theta_j \in \mathbb{R}$ $(j = 1, \cdots, N)$, $\beta \in \mathbb{R}$.



$$A_1 = \begin{bmatrix} w_1^T \\ \vdots \\ w_N^T \end{bmatrix}_{N \times d} \qquad b_1 = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_N \end{bmatrix}.$$

$$A_2 = [\alpha_1 \cdots \alpha_N]_{1 \times N}$$

$$b_2 = \beta \in \mathbb{R}^1.$$

$L = 2$, $N_0 = d$, $N_L = 1$, $W = N$.

Note: Since $\sigma \not\equiv$ const., by choosing all $w_j = 0$, we can absorb β. i.e., $NN(\sigma, 2, d, 1)$ consists of all G with $\beta = 0$.

Proof (Cybenko's proof)

Step 1 $\mathcal{N}_\sigma := NN(\sigma, L=2, d, 1)$ is a vector subspace of $C(K)$ by elementary constructions. (Or by direct verification.)

Step 2 If $f \in C(K) \setminus \overline{\mathcal{N}_\sigma}$ ($\overline{\mathcal{N}_\sigma}$ = the closure of $\mathcal{N}_\sigma$ in $C(K)$), then by the Hahn-Banach Thm, $\exists \varphi \in C(K)^*$ such that $\varphi(f) \neq 0$ and $\varphi = 0$ on $\overline{\mathcal{N}_\sigma}$.

By Riesz's Thm, $\exists \mu \in \mathcal{M}(K)$ — the space of Radon measures on K, s.t. $\varphi(g) = \int_K g \, d\mu \ \forall g \in C(K)$.

Step 3. For any $g(x) = \sigma(w^T x + \theta)$ ($x \in \mathbb{R}^d$), where $w \in \mathbb{R}^d$, $\theta \in \mathbb{R}$, we have $g \in \overline{\mathcal{N}_\sigma}$, hence $\varphi(g) = 0$, i.e.,

$$\varphi(g) = \int_K g \, d\mu = \int_K \sigma(w^T x + \theta) \, d\mu(x) = 0.$$

Extend μ to $\mathcal{M}(\mathbb{R}^d)$ (finite, signed Radon measures on $\mathbb{R}^d$) trivially, i.e., $\forall A \in \mathcal{B}_{\mathbb{R}^d}$ (all Borel sets of $\mathbb{R}^d$), $\mu(A) = \mu(A \cap K)$. Then $\mu \in \mathcal{M}_c(\mathbb{R}^d)$ (i.e., $\mu \in \mathcal{M}(\mathbb{R}^d)$ with compact support). Moreover

$$\int_{\mathbb{R}^d} \sigma(w \cdot x + \theta) \, d\mu(x) = 0 \quad \forall w \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}.$$

By a lemma (next lecture), the assumptions of σ lead to $\mu = 0$. Hence $\varphi(f) = \int_K f \, d\mu = 0$, a contradiction. (The continuity of σ is needed for $g(x) = \sigma(w^T x + \theta)$ to be in $C(K)$ so the $\varphi(g) = \int_K g \, d\mu$.)

QED

Remark The continuity of σ is needed in the theorem so that $NN(\sigma, 2, d, 1) \subseteq C(K)$.