# Mathematical Deep Learning Theory

Lec 1: Universal Approximation Theorem

Jinhee Paeng

Oct 5, 2023

Seoul National University

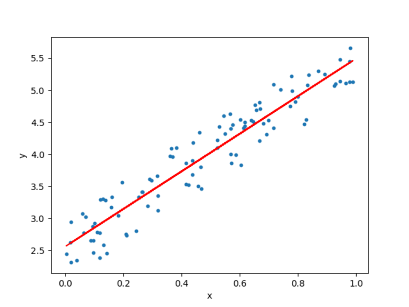
# Table of Contents

# Table of Contents

# Linear Regression

Consider, input $x \in \mathbb{R}^d$ and output $y \in \mathbb{R}$ following a model:

$$y = a^T x + b.$$




Traditionally, we estimate $a, b$ using *Least squares*.

# Features

We can perform *linear regression* multiple times:

$$y_i = a_i^T x + b_i,$$

and get a value of multiple features in return. For the simplicity,

$$y = Ax + b.$$

Now, the idea is predict the outcome using the generated features.

$$a^T(Ax + b) + b'.$$

Or, we can apply such idea repeatedly:

$$y = A^{(n)}(A^{(n-1)}(A^{(n-2)}(\cdots(A^{(1)}x + b^{(1)})\cdots + b^{(n-2)}) + b^{(n-1)}) + b^{(n)}.$$

# Activation function

**Problem.** Such model is equivalent to the linear regrassion.

$$y = Ax + b.$$

**Solution.** This is why we use nonlinear *activation function*:

$$\sigma : \mathbb{R} \to \mathbb{R}, \quad \sigma(x)_i := \sigma(x_i).$$

Some practical examples are *ReLU*, *Sigmoid*, *arctan*.

# Neural Network Structure

Using the activation function, we define a *Neural Network Structure* as:

$$y = A^{(n)}\sigma(A^{(n-1)}\sigma(\cdots\sigma(A^{(1)}x + b^{(1)})\cdots + b^{(n-1)}) + b^{(n)}.$$

We define *depth* of the *Neural Network* as $n$ value. We call each

$$A^{(i)}\cdot + b^{(i)}$$

as *layers*. The depth of Neural Network system is the number of layers.

Also, we call a *width* of Neural Network as the size of $A^{(i)}$ matrix.

**Example.** A $N$-width, 2-layer Neural Network can be written in form of

$$f_\theta(x) = \sum_{i=1}^{N} u_i\sigma(a_i^T x + b_i).$$

# Table of Contents

# 2-layer Neural Network

**Remark.** A $N$-width, 2-layer Neural Network can be written in form of

$$f_\theta(x) = \sum_{i=1}^{N} u_i \sigma(a_i^T x + b_i).$$

Here, $\theta$ is the parameter vector,

$$\theta = (u, a, b) \in \Theta_{(N)} = \mathbb{R}^{N+N\times d+N}.$$

**Goal.** The goal of this section is to prove 2-layer Neural Network system is *dense* subset of the set of continuous functions with $\|\cdot\|_\infty$.

# Table of Contents

# Universal Approximation Theorem

**Theorem (Universal Approximation Theorem)**

*Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying:*

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1.$$

*Let the domain $\Omega \subset \mathbb{R}^d$ be compact. Then the class of functions*

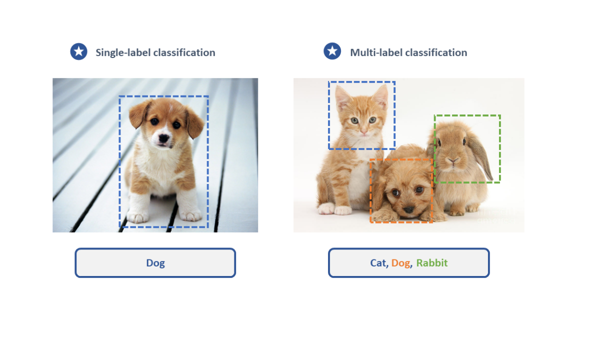
$$\bigcup_{N \in \mathbb{N}} \{f_\theta\}_{\theta \in \Theta_{(N)}} = \mathrm{span}\{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

*is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

# Universal Approximation Theorem

**Question.** Isn't compactness of $\Omega$ too strong?

**Answer.** It is sufficient. For example, consider a image classification.




In this case, the domain of input is $[0, 256]^{n\times m}$, a compact set.

# Proof of UAT

**Idea of Proof.**

The proof of Universal Approximation Theorem is done in two steps.

1. An activation function $\sigma$ that satisfies

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1$$

   is a *discriminatory* function.

2. When $\sigma$ is *discriminatory*, then

$$\overline{\mathrm{span}\{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}} = (\mathcal{C}(\Omega), \|\cdot\|_\infty).$$

□

# Discriminatory function

**Definition (Discriminatory function)**

A function $\sigma : \mathbb{R} \to \mathbb{R}$ is *discriminatory* if

$$\left[\forall a, b, \quad \int_{\Omega} \sigma(a^T x + b) d\mu(x) = 0\right] \Rightarrow \mu = 0,$$

for (finite signed regular Borel) measure $\mu \in \mathcal{M}(\Omega)$.

# Riesz Representation Theorem

**Theorem (Riesz-Markov-Kakutani Representation Theorem)**

*Let $\Omega \in \mathbb{R}^d$ be compact. Then for any bounded linear functional $L$ on $\mathcal{C}(\Omega)$, there is a unique signed regular Borel measure $\mu$ on $\Omega$ such that*

$$L[f] = \int_\Omega f(x) d\mu(x), \quad f \ in \in \mathcal{C}(\Omega).$$

**Remark.** When we write $L_\mu[f] := \int_\Omega f(x) d\mu(x)$, $\sigma$ is *discriminatory* if

$$\left[ L_\mu[\sigma(a^T \cdot + b)] = 0 \text{ for all } a, b \right] \Rightarrow L_\mu = 0.$$

# Part 1 of Proof

**Lemma**

*A function $\sigma$ that satisfies*

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1$$

*is a discriminatory function.*

# Part 1 of Proof

**Proof.**

1. Define $H_{a,b} := \{x : a^T x + b > 0\}$ and $\partial H_{a,b} := \{x : a^T x + b = 0\}$.
2. Define $\phi_{a,b}(x) := \sigma(a^T x + b)$. Then, $\phi_{\frac{a}{\delta},\frac{b}{\delta}} \xrightarrow{\delta \to 0} \gamma_t := \begin{cases} 1 & H_{a,b} \\ \sigma(t) & \partial H_{a,b} \\ 0 & o.w. \end{cases}$.
3. Since $\sigma$ is bounded, by Dominated convergence theorem,
$$L_\mu\left[\phi_{\frac{a}{\delta},\frac{b}{\delta}}\right] \xrightarrow{\delta \to 0} L_\mu[\gamma_t] = \mu(H_{a,b}) + \sigma(t)\mu(\partial H_{a,b}).$$
4. Suppose $L_\mu[\phi_{a,b}] = 0$ for all $a, b$.
   Then, $\mu(H_{a,b}) = \mu(\partial H_{a,b}) = 0$ since $\sigma$ is non-constant.
5. For any step function $s$, $L_\mu[s(a^T \cdot)] = 0$.
6. By DCT, $L_\mu[\sin(a^T \cdot)] = L_\mu[\cos(a^T \cdot)] = 0$.
7. $\mu = 0$ since its Fourier transform $\hat{\mu}(x) = \int_\Omega e^{ia^T x} d\mu(x) = 0$.

□

# Part 2 of Proof

**Lemma**

*When $\sigma$ is discriminatory, then*

$$\overline{S} = (\mathcal{C}(\Omega), \|\cdot\|_\infty),$$

*where*

$$S = \operatorname{span}\left\{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\right\}$$

# Part 2 of Proof

**Proof.**

Proof by contradiction. Suppose $\overline{S} = \mathcal{C}(\Omega)$ and $\exists g \in \mathcal{C}(\Omega) \setminus \overline{S}$.

1. Define a bounded linear functional $L : \overline{S} \bigoplus \mathrm{span}(g) \to \mathbb{R}$ as:

$$L[s + \lambda g] = \lambda, \quad s \in \overline{S}.$$

2. By Hahn-Banach Extension Theorem, extend $L$ to $\overline{L} : \mathcal{C}(\Omega) \to \mathbb{R}$.
3. By Riesz Representation Theorem, corresponding $\mu_{\overline{L}}$ exists.
4. Since $\overline{L} = 0$ on $\overline{S}$, we have $\overline{L}[\sigma(a^T x + b)] = 0$.
5. Since $\sigma$ is discriminatory, $\mu_{\overline{L}} = 0$. Thus, $\overline{L} = 0$ on $\mathcal{C}(\Omega)$.

Thus, $L[s + \lambda g] = \lambda \neq 0$ yields contradiction. □

# Extension to non-continuous function

**Remark.** However, most of the target functions of the given problem is not continuous. For example, in *classification*, $f_* : \Omega \to \{1, 2, \cdots, k\}$.

**Solution.** Lusin's theorem may solve such problem.

**Theorem (Lusin's Theorem)**

*Let $\Omega \subset \mathbb{R}^d$ be compact. Let $f : \Omega \to \mathbb{R}$ be a measurable function. For any $\epsilon > 0$, there exists a continuous function $f_\epsilon$ and $\Omega' \subseteq \Omega$ such that*

$$Vol(\Omega \setminus \Omega') < \epsilon, \qquad f(x) = f_\epsilon(x), \forall x \in \Omega'.$$

# Dense in $L^p$

We have established that $S^d$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.

**Question.** Can same be obtained on the Lebesgue $L^p$ space?

**Answer.** Sadly, no.

**Theorem**

*Let $d \geq 2$. For any Lebesgue measurable $\sigma$, any nonzero $g \in S^d$ satisfies*

$$\|g\|_{L^p} = \infty, \quad p \in [1, \infty).$$

# Dense in $L^p$

**Goal.** However, with finite nonnegative measure $\mu$, we can make $S^d$ dense in $L^p(\mu)$ for $p \in [1, \infty)$.

**Theorem**

*Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function satisfying:*

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1.$$

*Let the domain $\Omega \subset \mathbb{R}^d$ be compact. Then the class of functions*

$$S^d = \mathrm{span}\{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

*is dense in $L^p(\mu)$.*

# Table of Contents

# Generalization of UAT

**Remark.** The assumption used in *UAT* is quite strong.

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1.$$

It is not satisfied in widely used activation functions such as *ReLU*.

**Goal.** The result of *Universal Approximation Theorem* also holds for non-polynomial continuous $\sigma$.

# Stone-Weierstrass Theorem

**Theorem (Stone-Weierstrass Theorem)**

*Let $\Omega \subset \mathbb{R}^d$ be compact. Let $\mathcal{F} \subseteq (\mathcal{C}(\Omega), \|\cdot\|_\infty)$ be a subalgebra with nonzero constant function $c \in \mathcal{F}$. Then, $\mathcal{F}$ is dense if and only if*

$$\forall x, y \in \Omega \text{ with } x \neq y, \quad \exists f \in \mathcal{F} \text{ such that } f(x) \neq f(y).$$

# Stone-Weierstrass Theorem

Let's define a new set of functions:

$$g_\theta(x) = \sum_{i=0}^{N} u_i \prod_{j=1}^{M_i} \sigma(a_{ij}^T x + b_{ij}).$$

We can see that set of all function in form of $g_\theta$ forms an algebra.

**Corollary**

*The set of functions $\{g_\theta : \theta \in \mathbb{R}^*\}$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

**Remark.** $g_\theta$ is not a Neural Network form.

**Corollary**

*The set of functions $\bigcup_{N\in\mathbb{N}} \{f_\theta\}_{\theta\in\Theta_{(N)}}$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$ if $\sigma = \sin$.*

## Generalization of UAT

**Goal.** The goal is to prove in general case, when $\sigma$ is non-polynomial.

**Theorem**

*Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-polynomial continuous function.*
*Let the domain $\Omega \subset \mathbb{R}^d$ be compact. Then the class of functions*

$$S^d = \bigcup_{N \in \mathbb{N}} \{f_\theta\}_{\theta \in \Theta_{(N)}} = \mathrm{span}\{\sigma(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

*is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$.*

# Reduction to 1-dimension

**Idea.** First let's simplify the problem into 1-dimension.

**Lemma**

*Let $\sigma \in \mathcal{C}(\mathbb{R})$ makes $S^1$ dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$ for any compact $K \subset \mathbb{R}$. Then, $S^d$ is dense in $(\mathcal{C}(\Omega), \|\cdot\|_\infty)$ for any compact $\Omega \subset \mathbb{R}^d$.*

# Reduction to 1-dimension

**Proof.**

Choose any target function $f_\star \in \mathcal{C}(\Omega)$.

1. Since $\mathrm{span}\{\sin(a^T x + b) : a \in \mathbb{R}^d, b \in \mathbb{R}\}$ is dense in $\Omega \subset \mathbb{R}^d$,

$$\left| f_\star(x) - \sum_{i=1}^{N} u_i \sin(a_i^T x + b_i) \right| < \frac{\epsilon}{2}, \quad \forall x \in \Omega.$$

2. Let $D := \sup_{x \in \Omega, i \in [N]} |a_i^T x|$. Since $S^1$ is dense in $\mathcal{C}([-D, D])$,

$$\left| u_i \sin(a_i^T x + b_i) - f_{\theta_i}(a_i^T x) \right| \le \epsilon / 2N$$

3. Thus, there exists $f_\theta \in S^d$ such that $|f_\star(x) - f_\theta(x)| < \epsilon$ for all $x \in \Omega$.

□

# Case: $\sigma \in \mathcal{C}^\infty(\mathbb{R})$

Let's begin with the simple case where $\sigma \in \mathcal{C}^\infty(\mathbb{R})$.

**Lemma**

*Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-polynomial $\mathcal{C}^\infty(\mathbb{R})$ function.*
*Let the domain $K \subset \mathbb{R}$ be compact. Then the class of functions*

$$S^1 = \mathrm{span}\{\sigma(ax + b) : a \in \mathbb{R}, b \in \mathbb{R}\}.$$

*is dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$.*

# Case: $\sigma \in \mathcal{C}^\infty(\mathbb{R})$

**Proof.**

1. For all $t \in \mathbb{R}$, $x\sigma'(t) \in \overline{S^1}$ since the compactness of $K$ gives

$$x\sigma'(t) = \frac{d}{ds}\sigma(xs+t)\bigg|_{s=0} = \lim_{h\to 0}\frac{\sigma(xh+t)-\sigma(t)}{h} \in \overline{S^1}.$$

2. Similarly, for all $k \in \mathbb{N}$ and $t \in \mathbb{R}$, $x^k\sigma^{(k)}(t) \in \overline{S^1}$.
3. Since $\sigma$ is non-polynomial, there exists $t$ that $\sigma^{(k)}(t) \neq 0$. Thus,

$$x^k \in \overline{S^1}, \quad \forall k \in \mathbb{N}.$$

4. From S-W Thm, $\mathrm{span}\{x^k : k \in \mathbb{N}\}$ is dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$.
5. Since $\mathrm{span}\{x^k : k \in \mathbb{N}\} \subseteq \overline{S^1}$, $S^1$ is dense in $(\mathcal{C}(K), \|\cdot\|_\infty)$.

□

# Generalization of UAT

We will use the result with $\sigma \in \mathcal{C}^\infty(\mathbb{R})$ assumption using the *mollifier* $\phi_\delta$:

$$\phi_\delta := \frac{1}{\delta \int_{\mathbb{R}} \Psi(t)dt} \Psi(t/\delta), \quad \Psi(t) := \begin{cases} \exp\left(-\frac{1}{1-t^2}\right) & t \in (-1,1) \\ 0 & \text{otherwise} \end{cases}.$$

When the continuous function $\sigma$ is given, define $\mathcal{C}^\infty(\mathbb{R})$ function $\sigma_\delta$ as:

$$\sigma_\delta(r) := \int_{\mathbb{R}} \sigma(r-t)\phi_\delta(t)dt \in \mathcal{C}^\infty(\mathbb{R}).$$

We can check that for a compact $K \subset \mathbb{R}$, $\sigma_\delta$ is close to $\sigma$ with small $\delta$:

$$\lim_{\delta \to 0} \left[ \sup_{r \in K} |\sigma_\delta(r) - \sigma(r)| \right] = 0$$

# Generalization of UAT

**Proof.**

We can show following two facts.

- $\sigma_\delta \in \overline{S^1} = \overline{\mathrm{span}\{\sigma(ax+b) : a \in \mathbb{R}, b \in \mathbb{R}\}}$.
  (It can be shown using Riemann sum of $\sigma_\delta(r) = \int_{\mathbb{R}} \sigma(r-t)\phi_\delta(t)dt$.)
- Since $\sigma$ is non-polynomial, for each $k \in \mathbb{N}$ there exist $\delta > 0$ such that $\sigma_\delta$ is not a polynomial of degree at most $k$.
  (Set of polynomials of degree at most $k$ is closed set, and $\sigma_\delta \overset{\delta \to 0}{\to} \sigma$.)

This gives

$$\mathrm{span}\{x^k : k \in \mathbb{N}\} \subseteq \bigcup_{\delta > 0} \overline{\mathrm{span}\{\sigma_\delta(sr+t) : s, t \in \mathbb{R}\}} \subseteq \overline{S^1}$$

since $x^k \sigma^{(k)}(t) \in \overline{S^1}$ and we can make $\sigma^{(k)}(t) \neq 0$.
Finally, by S-W Thm, $\overline{\mathrm{span}\{x^k : k \in \mathbb{N}\}} = \mathcal{C}(K)$ concludes the proof. □

# Table of Contents

# Conclusion

Throughout this lecture we proved 2-layer Neural Network structure:

$$f_\theta(x) = \sum_{i=1}^{N} u_i \sigma(a_i^T x + b_i)$$

forms a *dense* subset of the set of continuous function. This gives the mathematical foundation of why neural network structure may approximate the target function well.

**Next lecture.** In the next lecture we will quantify the approximation capability. We will show that (in 2-layer NN) the error can be controlled in the scale of $\mathcal{O}(1/N)$, the inverse of width.