GT Deep Learning: 05

Vincent Perrollaz

10 Mars 2026

Statement

Base Case n=m=1

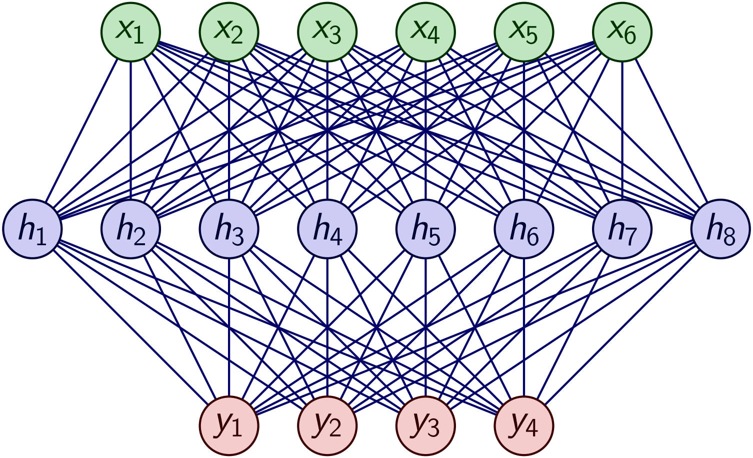
Mixed Case m=1

Weierstrass Theorem

Conclusion

One Layer Perceptron:

input $\mathbb{R}^6$, hidden layers $\mathbb{R}^8$, output $\mathbb{R}^4$



Formula Previous Case

$$1 \leq k \leq 8, \quad h_k = \sigma\left(\sum_{j=1}^{6} w_{1,k,j} x_j + b_{1,k}\right) \tag{1}$$

$$1 \leq i \leq 4, \quad y_i = \sum_{k=1}^{8} w_{2,i,k} h_k + b_{2,i} \tag{2}$$

Formula General Case

n inputs, d hidden neurons, m outputs

$$1 \leq i \leq m, \quad y_i = \sum_{k=1}^{d} w_{2,i,k}\sigma(w_{1,k} \cdot x + b_{2,k}). \quad (3)$$

$\mathcal{N}(\sigma, \mathbb{R}^n, \mathbb{R}^m)$ space of such functions

Statement

Cybenko 1989, Hornik-Stinchcombe-White 1989.

- $\forall n, m \geq 1$
- $\forall X > 0$
- $\forall \sigma \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$

$$\mathcal{N}(\sigma, \mathbb{R}^n, \mathbb{R}^m) \text{ dense in } (\mathcal{C}^0([-X, X]^n, \mathbb{R}^m), \| \cdot \|_\infty) \iff \sigma \text{ not polynomial} \quad (4)$$

Base Case I

- Equivalent to show

$$R(\sigma) := \mathrm{Vect}(x \mapsto \sigma(wx + b) \,:\, (w, b) \in \mathbb{R}^2) \quad (5)$$
$$\implies \overline{R(\sigma)} = \mathcal{C}^0([-X, X]; \mathbb{R}). \quad (6)$$

- Fix w, b then

$$x \mapsto \frac{\sigma((w+t)x + b) - \sigma(wx + b)}{t} \in R(\sigma) \quad (7)$$

- Letting $t \to 0^+ \implies$

$$\{x \mapsto x\sigma'(wx + b) \,:\, (w, b) \in \mathbb{R}^2\} \subset \overline{R(\sigma)}. \quad (8)$$

Base Case II

- Recurrence $\implies$

$$\{x \mapsto x^n \sigma^{(n)}(wx+b) \ : \ (n,w,b) \in \mathbb{N} \times \mathbb{R}^2\} \subset \overline{R(\sigma)}. \tag{9}$$

- σ non polynomial $\implies$

$$\forall n \geq 0, \quad \exists z \in \mathbb{R}, \quad \sigma^{(n)}(z) \neq 0. \tag{10}$$

- Therefore

$$\mathbb{R}[X] \subset \overline{R(\sigma)} \tag{11}$$

- Conclusion by Weierstrass theorem.

Mixed Case I

- Equivalent to show

$$R_n(\sigma) := \mathrm{Vect}(x \mapsto \sigma(w \cdot x + b) \ : \ (w, b) \in \mathbb{R}^n \times \mathbb{R})$$
$$\implies \overline{R_n(\sigma)} = \mathcal{C}^0([-X, X]^n; \mathbb{R}).$$

- Fix w, b then

$$x \mapsto \frac{\sigma((w + te_i)x + b) - \sigma(wx + b)}{t} \in R_n(\sigma)$$

- Letting $t \to 0^+ \implies$

$$\{x \mapsto x_i \sigma'(wx + b) \ : \ (w, b) \in \mathbb{R}^n \times \mathbb{R}\} \subset \overline{R_n(\sigma)}$$

Mixed Case II

- Recurrence $\implies$

$$\{x \mapsto (\Pi_{i=1}^{n} x_i^{d_i})\sigma^{(d_1+\cdots+d_n)}(w\cdot x + b) \\ : (d_1,\ldots,d_n,w,b) \in \mathbb{N}^n \times \mathbb{R}^n \times \mathbb{R}\} \subset \overline{R_n(\sigma)}.$$

- σ non polynomial $\implies$

$$\forall n \geq 0, \quad \exists z \in \mathbb{R}, \quad \sigma^{(n)}(z) \neq 0. \tag{12}$$

- Therefore

$$\mathbb{R}[X_1,\ldots,X_n] \subset \overline{R_n(\sigma)} \tag{13}$$

- Conclusion by Weierstrass theorem.

General Case

- Edge weight $= 0 \iff$ No edge!
- To fuse two neural networks
  1. Juxtapose hidden neurons
  2. Cancel the weight of the artificial edges!
- Blackboard: Fuse two $1/2/1$ networks

Statement

Base Case n=m=1

Mixed Case m=1

Weierstrass Theorem

Conclusion

Weierstrass Theorem I

- $u \in \mathcal{C}^0(\mathbb{R}^n; \mathbb{R})$ polynomial growth at ∞
- $\epsilon > 0$

$$u_\epsilon := x \mapsto \int_{\mathbb{R}^n} u(x-y) e^{-\pi \|y\|^2/\epsilon} \mathrm{d}y_1 \ldots \mathrm{d}y_n / \epsilon^{n/2}$$

- u_ϵ entire function!
- $\epsilon \to 0 \implies u_\epsilon \to u$ locally uniformly
- u more regular $\implies$ Stronger CV!

Weierstrass Theorem II

- u_ϵ entire $\implies$ truncate power series
- K compact $\implies$ $\overline{\mathbb{R}(X_1, \ldots, X_n)} = \mathcal{C}^0(K, \mathbb{R})$
- In fact for Frechet topologies ⚠
- u defined in K compact $\implies$

$$\tilde{u} := x \in \mathbb{R}^n \mapsto \begin{cases} u(x) & \text{if } x \in K \\ \inf_{y \in K} u(y) + \frac{\mathrm{d}(x,y)}{\mathrm{d}(x,K)} - 1 & \text{else} \end{cases}$$

- $\tilde{u}_{|K} = u$, $\tilde{u}$ continuous, linear growth

Questions/Remarks

- Initiated by Kolmogorov (1957), see KAN 13th Hilbert Problem!
- Other Proofs? Other activation function? ReLU, Heaviside, exp, sin!
- Approximation speed? Sparsity? Architecture impact? Number of layers? Barron's Theorem!
- More General Approximation Theorem? Probability Distribution, Neural Operators!