

Deep Learning and Numerical PDEs

Shallow Neural Network Approximation

Jinchao Xu

KAUST and Penn State

xu@multigrid.org

Morgan State University, June 23, 2023

CBMS Lecture Series

NSF DMS-2111387, KAUST Baseline Research Fund

- 1 Shallow neural networks
- 2 Dictionary and variation spaces
- 3 Approximation properties of shallow neural networks
- 4 Metric Entropy
- 5 Summary

Shallow Neural Networks

$$\Sigma_n^\sigma = \left\{ \sum_{i=1}^n a_i \sigma(w_i \cdot x + b_i), w_i \in \mathbb{R}^d, b_i \in \mathbb{R} \right\} \quad (1)$$

Common activation functions:

- Heaviside $\sigma = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$
- Sigmoidal $\sigma = (1 + e^{-x})^{-1}$
- Rectified Linear $\sigma = \max(0, x)$
- Power of a ReLU $\sigma = [\max(0, x)]^k$
- Cosine $\sigma = \cos(x)$
- ...

How efficient is Σ_n^σ for approximation?

Approximation Rates for Shallow Neural Networks

Spectral Barron Space:

$$\|f\|_{\mathcal{B}^s} := \int_{\mathbb{R}^d} (1 + |\omega|)^s |\hat{f}(\omega)| d\omega \quad (2)$$

- Defined on domains via minimal extensions

Approximation Rate:

Theorem (Barron 1993)

For sigmoidal activation functions σ and bounded domain Ω ,

$$\inf_{u_N \in \Sigma_N^\sigma} \|u - u_N\|_{L^2(\Omega)} \lesssim N^{-\frac{1}{2}} \|u\|_{\mathcal{B}^1}. \quad (3)$$

Extensions:

- Compactly supported activation functions
- Smooth activation functions
- etc.

Ref: H. Mhaskar, C. Micchelli 1992, M. Leshno, V. Lin, A. Pinkus and S. Schocken 1993; K.Hornik, M.Stinchcombe, H.White and P.Auer 1994

Approximation Rates for Shallow Neural Networks

Our results extend these rates to larger classes of activation functions:

Theorem (Siegel and X 2020)

For activation functions $\sigma \in W_{\text{local}}^{m,\infty}$ with polynomial decay and bounded domains Ω ,

$$\inf_{u_N \in \Sigma_N^\sigma} \|u - u_N\|_{H^m(\Omega)} \lesssim N^{-\frac{1}{2}} \|u\|_{\mathcal{B}^{m+1}}. \quad (4)$$

With a somewhat worse rate of decay, even (almost) all activation functions:

Theorem (Siegel and X 2020)

Suppose that $\sigma \in L^\infty$ and $\hat{\sigma}$ (as a distribution) is a non-zero bounded function on some open interval I , then

$$\inf_{u_N \in \Sigma_N^\sigma} \|u - u_N\|_{L^2(\Omega)} \lesssim N^{-\frac{1}{4}} \|u\|_{\mathcal{B}^1}. \quad (5)$$

Ex:

- $\sigma \in BV(\mathbb{R})$
- $\sigma \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$

Ref: Siegel and Xu 2020

Approximation Rates for Shallow Neural Networks

Our results improve this for ReLU^k activation functions

Theorem (Siegel and X 2022)

Suppose that $\sigma = \max(0, x)^k$. Then we have

$$\inf_{u_N \in \Sigma_N^\sigma} \|u - u_N\|_{L^2(\Omega)} \lesssim N^{-\frac{1}{2}} \|u\|_{\mathcal{B}^{\frac{1}{2}}}. \quad (6)$$

- less smoothness required

Theorem (Siegel and X 2022)

Suppose that $\sigma = \max(0, x)^k$ and $s \geq (d+1)(k+1/2) + 1/2$. Then we have

$$\inf_{u_N \in \Sigma_N^\sigma} \|u - u_N\|_{L^2(\Omega)} \lesssim N^{-(k+1)} \log(N) \|u\|_{\mathcal{B}^s}. \quad (7)$$

- More smoothness gives better rates

Ref: Siegel and Xu 2022

- 1 Shallow neural networks
- 2 Dictionary and variation spaces**
- 3 Approximation properties of shallow neural networks
- 4 Metric Entropy
- 5 Summary

Perspective: Dictionary Approximation

$\mathbb{D} \subset X$ for a Banach space X is a dictionary if

- \mathbb{D} is bounded, i.e. $|\mathbb{D}| = \sup_{d \in \mathbb{D}} \|d\|_X < \infty$
- \mathbb{D} is symmetric, i.e. $d \in \mathbb{D} \rightarrow -d \in \mathbb{D}$

Non-linear dictionary approximation:

$$\Sigma_n(\mathbb{D}) := \left\{ \sum_{i=1}^n a_i d_i, d_i \in \mathbb{D} \right\} \quad (8)$$

Stable dictionary approximation:

$$\Sigma_n^M(\mathbb{D}) := \left\{ \sum_{i=1}^n a_i d_i, d_i \in \mathbb{D}, \sum_{i=1}^n |a_i| \leq M \right\} \quad (9)$$

Ref: Siegel, J. W. & Xu, J. (2023)

Variation spaces

- Take

$$B_1(\mathbb{D}) := \overline{\text{conv}(\mathbb{D})} = \overline{\left\{ \sum_{i=1}^n a_i d_i : \sum_{i=1}^n |a_i| \leq 1, n \in \mathbb{N} \right\}} \quad (10)$$

- Define $\mathcal{K}_1(\mathbb{D})$ -norm by

$$\|f\|_{\mathcal{K}_1(\mathbb{D})} := \inf \{ r > 0 : f \in B_1(\mathbb{D}) \} = \inf \left\{ \sum_{i=1}^n |a_i| : f = \sum_{i=1}^n a_i h_i \right\}.$$

Clearly, the unit ball of $\mathcal{K}_1(\mathbb{D})$ is $B_1(\mathbb{D})$.

- $\{f \in X : \|f\|_{\mathcal{K}_1(\mathbb{D})} \leq \infty\}$ is a Banach space

Ref: DeVore (1998), Siegel, J. W. & Xu, J. (2023)

Neural Network Dictionaries with Activation Function

- What is the relationship with shallow neural networks?
- Given an activation function σ and domain $\Omega \subset \mathbb{R}^d$, consider the dictionary

$$\mathbb{D}_\sigma^d = \{\sigma(\omega \cdot x + b), \omega \in \mathbb{R}^d, b \in \mathbb{R}\} \subset L^p(\Omega) \quad (11)$$

- For some σ , may need to restrict ω and b to ensure boundedness
- In this case

$$\Sigma_n(\mathbb{D}_\sigma^d) = \left\{ \sum_{i=1}^n a_i \sigma(\omega_i \cdot x + b_i) \right\} \quad (12)$$

is exactly the set of shallow neural networks with width n

- Typical σ : ReLU^k activation functions.

ReLU^k Activation Function

- Consider the ReLU^k activation function

$$\sigma_k(x) = \begin{cases} 0 & x \leq 0 \\ x^k & x > 0. \end{cases} \quad (13)$$

- In this case, $\sigma_k(\omega \cdot x + b)$ is not uniformly bounded in $L^p(\Omega)$!
- Must restrict ω and b , so consider the dictionary

$$\mathbb{P}_k^d = \{\sigma_k(\omega \cdot x + b), \omega \in S^{d-1}, b \in [-2, 2]\} \subset L^2(B_1^d), \quad (14)$$

where B_1^d is the unit ball in \mathbb{R}^d .

$\mathcal{K}_1(\mathbb{P}_k^d)$ is the variation space corresponding to shallow ReLU^k networks

Integral Representations of $\|f\|_{\mathcal{K}_1(\mathbb{D})}$

- If $\mathbb{D} \subset X$ is dense, the norm $\|f\|_{\mathcal{K}_1(\mathbb{D})} := \inf\{r > 0 : f \in B_1(\mathbb{D})\}$ can be written equivalently as

$$\begin{aligned}\|f\|_{\mathcal{K}_1(\mathbb{D})} &= \inf \left\{ \sum_{i=1}^n |a_i| : f = \sum_{i=1}^n a_i h_i \right\} \\ &= \inf \left\{ \int_{\mathbb{D}} d|\mu| : f = \int_{\mathbb{D}} h d\mu \right\}.\end{aligned}$$

- For ReLU^k neural network dictionaries, we can write

$$\|f\|_{\mathcal{K}_1(\mathbb{D}_\sigma^d)} = \inf_{\mu \in \mathcal{B}(\mathbb{S}^d \times [-2, 2])} \left\{ \int_{\mathbb{S}^d \times [-2, 2]} d|\mu| : f = \int_{\mathbb{S}^d \times [-2, 2]} \sigma(w \cdot x + b) d\mu(w, b) \right\},$$

where $\mathcal{B}(\mathbb{S}^d \times [-2, 2])$ is the set of Borel measures on $\mathbb{S}^d \times [-2, 2]$.

Ref: E, W (2017), Siegel, J. W. & Xu, J. (2023)

What is $\mathcal{K}_1(\mathbb{P}_k^d)$? ($d = 1$)

In this case, $\mathbb{P}_k^d = \{(\pm x - b)_+^k : b \in [-2, 2]\}$. We claim

$$\|f\|_{\mathcal{K}_1(\mathbb{P}_k^1)} \sim \|f\|_{L_\infty([-1, 1])} + \|f^{(k)}\|_{BV[-1, 1]}.$$

Proof: By Peano Kernel Formula, on $[-1, 1]$,

$$\begin{aligned} f(x) &= f(-1) + f^{(1)}(-1)(x+1) + \frac{f^{(2)}(-1)}{2}(x+1)^2 + \cdots + \frac{f^{(k)}(-1)}{k!}(x+1)^k + \int_{-1}^x \frac{f^{(k+1)}(y)}{(k+1)!}(x-y)^k dy \\ &= f(-1) + f^{(1)}(-1)(x+1) + \frac{f^{(2)}(-1)}{2}(x+1)^2 + \cdots + \frac{f^{(k)}(-1)}{k!}(x+1)^k + \int_{-1}^1 \frac{f^{(k+1)}(y)}{(k+1)!}(x-y)_+^k dy \end{aligned}$$

The last gives an integral representation if $f^{(k+1)} \in L_1([0, 1])$. Since each polynomial of degree $j \leq k$ can be recovered from polynomials of type $(x+b)_+^k$, we can represent $(x+1), \dots, (x+1)^k$ be finite linear combinations of elements in \mathbb{P}_k^d . This shows

$$\|f\|_{\mathcal{K}_1(\mathbb{P}_k^1)} \lesssim \sum_{1 \leq j \leq k} \|f^{(j)}\|_{L_\infty([-1, 1])} + \|f^{(k)}\|_{BV[-1, 1]} \lesssim \|f\|_{L_\infty([-1, 1])} + \|f^{(k)}\|_{BV[-1, 1]}$$

The other direction is obvious by definition.

What is $\mathcal{K}_1(\mathbb{P}_k^d)$? ($d > 1$)

Use the Radon transform on \mathbb{R}^d . Given $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the Radon transform is

$$\mathcal{R}f(w, b) := \int_{w \cdot x + b = 0} f(x) dS(x),$$

where S is the natural hypersurface measure.

Suppose $f \in C_c^\infty(\mathbb{R}^d)$, we will reconstruct f from $\mathcal{R}f$.

Fix $w \in \mathbb{S}^{d-1}$, consider the univariate Fourier transform \mathcal{F} on the variable b , we have

$$\mathcal{F}\mathcal{R}f(w, t) = \int_{\mathbb{R}} e^{-2\pi i t b} \int_{w \cdot x + b = 0} f(x) dS(x) db = \int_{\mathbb{R}^d} e^{2\pi i t w \cdot x} f(x) dx = \hat{f}(-tw).$$

So we can reconstruct f from the Radon transform using the Fourier transform:

$$\begin{aligned} f(x) &= \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i \xi \cdot x} d\xi = \int_{\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} \hat{f}(tw) e^{2\pi i t w \cdot x} |t|^{d-1} dt dw \\ &= \int_{\mathbb{S}^{d-1}} \int_{-\infty}^{\infty} \mathcal{F}\mathcal{R}f(w, -t) e^{2\pi i t w \cdot x} |t|^{d-1} dt dw = \int_{\mathbb{S}^{d-1}} \tilde{\mathcal{R}}f(w, -w \cdot x) dw, \end{aligned}$$

where

$$\tilde{\mathcal{R}}f(w, b) = \mathcal{F}^{-1} \left[|t|^{d-1} \mathcal{F}\mathcal{R}f(w, t) \right] (b).$$

What is $\mathcal{K}_1(\mathbb{P}_k^d)$? ($d > 1$)

Consider the Fourier transform for functions in the real space. Using the basic property of univariate Fourier transform, if d is odd,

$$\tilde{\mathcal{R}}f(w, b) = (-i)^{d-1} \left(\frac{\partial}{\partial b} \right)^{d-1} \mathcal{R}f(w, b).$$

If d is even, notice that $g(t) = \frac{i}{\pi t}$ is the Fourier transform of $\text{sgn}(x)$, we have

$$\tilde{\mathcal{R}}f(w, b) = p.v. \int_{-\infty}^{\infty} \frac{i}{\pi(b-t)} (-i)^{d-1} \left(\frac{\partial}{\partial b} \right)^{d-1} \mathcal{R}f(w, b) dt.$$

In this case, $\tilde{\mathcal{R}}f$ is the Hilbert transform of $\left(\frac{\partial}{\partial b} \right)^{d-1} \mathcal{R}f(w, b)$ multiplied with i .

Now we use

$$\left\| \left(\frac{d}{dt} \right)^{k+d-1} \mathcal{R}f \right\|_{BV(dt)} < \infty, \quad d \text{ is odd}, \quad \left\| \mathcal{H} \left(\frac{d}{dt} \right)^{k+d-1} \mathcal{R}f \right\|_{BV(dt)} < \infty, \quad d \text{ is even}.$$

Then

$$\|f\|_{\mathcal{K}_1(\mathbb{P}_k^d)} \lesssim \begin{cases} \int_{\mathbb{S}^{d-1}} \left\| \left(\frac{d}{dt} \right)^{k+d-1} \mathcal{R}f \right\|_{BV(dt)} dw, & d \text{ is odd}, \\ \int_{\mathbb{S}^{d-1}} \left\| \mathcal{H} \left(\frac{d}{dt} \right)^{k+d-1} \mathcal{R}f \right\|_{BV(dt)} dw, & d \text{ is even}. \end{cases} \quad (15)$$

The Spectral Barron Space

- Let $\Omega = \{x \in \mathbb{R}^d : |x| \leq 1\}$ and consider the dictionary

$$\mathbb{D} = \mathbb{F}_s^d := \{(1 + |\omega|)^{-s} e^{2\pi i \omega \cdot x} : \omega \in \mathbb{R}^d\}. \quad (16)$$

- The spectral Barron norm is characterized by

$$\|f\|_{\mathcal{B}^s} \simeq \|f\|_{\mathcal{K}_1(\mathbb{F}_s^d)} \quad (17)$$

- Property:

$$H^{s+\frac{d}{2}+\varepsilon}(\Omega) \hookrightarrow \mathcal{B}^s(\Omega) \hookrightarrow W^{s,\infty}(\Omega). \quad (18)$$

Ref: Siegel, J. W. & Xu, J. (2023)

- 1 Shallow neural networks
- 2 Dictionary and variation spaces
- 3 Approximation properties of shallow neural networks**
- 4 Metric Entropy
- 5 Summary

Stable neural network and approximation properties

$$\Sigma_{n,M}^\sigma := \left\{ \sum_{i=1}^n a_i h_i, h_i \in \mathbb{D}_\sigma, \sum_{i=1}^n |a_i| \leq M \right\} \quad (19)$$

Theorem (Siegel & Xu, 2021-2022)

A function $u \in L^2(\Omega)$ can be approximated at all, i.e.

$$\lim_{n \rightarrow \infty} \inf_{u_n \in \Sigma_{n,M}^\sigma} \|u - u_n\|_{L^2(\Omega)} = 0, \quad (20)$$

for some $M > 0$ with $\sigma \in L^\infty(\mathbb{R})$, if and only if $u \in \mathcal{K}_1(\mathbb{D}_\sigma)$. Furthermore,

$$\inf_{u_n \in \Sigma_{n,M}^\sigma} \|u - u_n\|_{L^2(\Omega)} \leq C n^{-\frac{1}{2}} \|u\|_{\mathcal{K}_1(\mathbb{D}_\sigma)}. \quad (21)$$

If $\sigma = \text{ReLU}^k$,

$$\inf_{u_n \in \Sigma_{n,M}^\sigma} \|u - u_n\|_{L^2(\Omega)} \leq C n^{-\frac{1}{2} - \frac{2k+1}{2d}} \|u\|_{\mathcal{K}_1(\mathbb{P}_k^d)}. \quad (22)$$

- Earlier results: Barron, A. R. (1993), Makovoz, Y. (1996), Klusowski, J. M. & Barron, A. R. (2018), E, W., Ma, C. & Wu, L. (2019), Xu, J. (2021), Siegel, J. W. & Xu J. (2021)

Abstract Dictionary Approximation for Variation Spaces

What rates can be obtained on $\mathcal{K}_1(\mathbb{D})$ for $\Sigma_n(\mathbb{D})$?

Theorem (Barron, Jones, Maurey)

In a Hilbert space, we always have the approximation rate

$$\inf_{f_n \in \Sigma_n(\mathbb{D})} \|f - f_n\|_H \leq |\mathbb{D}| \|f\|_{\mathcal{K}_1(\mathbb{D})} n^{-\frac{1}{2}}. \quad (23)$$

- We actually have $f_n \in \Sigma_n^M(\mathbb{D})$ for $M = \|f\|_{\mathcal{K}_1(\mathbb{D})}$
- Also holds more generally in type-2 Banach spaces
- E.g. in L^p for $2 \leq p < \infty$
- This theorem can be proved using the sampling argument or greedy algorithm

Optimal in the worst case over all \mathbb{D} : Consider the dictionary $\mathbb{D} = \{e_1, e_2, \dots\} \subset \ell^2(\mathbb{N})$. Then

$$\|f\|_{\mathcal{K}_1(\mathbb{D})} = \|f\|_{\ell^1} = \sum_{j=1}^{\infty} |f_j|.$$

Given any $n \in \mathbb{N}$, take $f = \frac{1}{2n} \sum_{j=1}^{2n} e_j \in B_1(\mathbb{D})$. Then for any $f_n \in \Sigma_n(\mathbb{D})$,

$$\|f - f_n\|_{\ell^2}^2 \geq \frac{1}{4n^2} \sum_{j=1}^n 1 = \frac{1}{4n}.$$

Ref: Pisier (1983), Jones (1992), Barron (1993)

Sampling argument

- ① Let $f \in B_1(D)$, for any $\epsilon > 0$, there exist ρ_i, h_i with $i = 1, \dots, N$, such that

$$\|f - g\|_H \leq \epsilon, \quad \text{with} \quad g = \sum_{i=1}^N a_i h_i, \quad \text{and} \quad \sum_{i=1}^N a_i = 1. \quad (24)$$

Without loss of generality, assume $a_i \geq 0$.

- ② For any g_{i_1, \dots, i_n} , define

$$\mathbb{E}_n g_{i_1, \dots, i_n} := \sum_{i_1, \dots, i_n=1}^N g_{i_1, \dots, i_n} \prod_{j=1}^n a_{i_j}$$

- ③ For $g_{i_1, \dots, i_n} = \frac{1}{n} \sum_{j=1}^n h_{i_j}$,

$$\mathbb{E}_n \|g - g_{i_1, \dots, i_n}\|_H^2 = \frac{1}{n} \left(\mathbb{E}(\|h\|_H^2) - (\mathbb{E}\|h\|_H)^2 \right) \leq \frac{1}{n} \mathbb{E}(\|h\|_H^2) \leq \frac{1}{n} \|\mathbb{D}\|^2.$$

- ④ There exist $\{i_j^*\}$ such that

$$\|g - g_{i_1^*, \dots, i_n^*}\|_H \leq n^{-\frac{1}{2}} \|\mathbb{D}\|.$$

- ⑤ Let $g_n = \frac{1}{n} \sum_{j=1}^n h_{i_j^*}$. Then,

$$\|f - g_n\|_H \leq \|f - g\|_H + \|g - g_n\|_H \leq \epsilon + n^{-\frac{1}{2}} \|\mathbb{D}\|.$$

Relaxed Greedy Algorithm (Jones 1992)

1 Let $\|f\|_{\mathcal{K}_1(\mathbb{D})} \leq 1$ and consider the *relaxed greedy algorithm*

$$f_1 = 0, \quad h_n = \arg \max_{h \in \mathbb{D}} \langle f - f_{n-1}, h \rangle, \quad f_n = \left(1 - \frac{1}{n}\right) f_{n-1} + \frac{1}{n} h_n \quad (25)$$

► Note that $f_n \in \Sigma_{n,1}(\mathbb{D})$

2 Claim: $\|f - f_n\| \leq 2|\mathbb{D}|n^{-\frac{1}{2}}$

3 Proof:

► We only need prove this for those $f \in B_1(\mathbb{D})$ that can be written as $f = \sum_{i=1}^n a_i g_i$, $a_i \geq 0$, $g_i \in \mathbb{D}$, $\sum_{i=1}^n a_i \leq 1$.

► Note that $\|f - f_n\|^2 = \left\| \left(1 - \frac{1}{n}\right) (f - f_{n-1}) + \frac{1}{n} (f - h_n) \right\|^2$, expand:

$$\|f - f_n\|^2 = \left(1 - \frac{1}{n}\right)^2 \|f - f_{n-1}\|^2 + \frac{2}{n} \left(1 - \frac{1}{n}\right)^2 \langle f - f_{n-1}, f - h_n \rangle + \frac{1}{n^2} \|f - h_n\|^2 \quad (26)$$

► By the argmax property: $\langle f - f_{n-1}, h_n \rangle \geq \sum_{i=1}^n a_i \langle f - f_{n-1}, g_i \rangle = \langle f - f_{n-1}, f \rangle$

► By boundedness of \mathbb{D} : $\|f - h_n\|^2 \leq 4|\mathbb{D}|^2$

► Get

$$\|f - f_n\|^2 \leq \left(1 - \frac{1}{n}\right)^2 \|f - f_{n-1}\|^2 + \frac{4|\mathbb{D}|^2}{n^2}$$

► Base case: $\|f - f_1\|^2 \leq |\mathbb{D}|^2 \leq 4|\mathbb{D}|^2$. Induction gives

$$\|f - f_n\|^2 \leq \left[\left(1 - \frac{1}{n}\right)^2 \frac{1}{n-1} + \frac{1}{n^2} \right] 4|\mathbb{D}|^2 = \frac{1}{n} 4|\mathbb{D}|^2.$$

Improving the Rates

Previous results of $n^{-\frac{1}{2}}$:

- Optimal in general
- Can be improved for certain specific \mathbb{D}

Theorem (Makovoz)

Consider the Heaviside activation function with dictionary \mathbb{P}_0^d . Then we have

$$\inf_{f_n \in \Sigma_n(\mathbb{P}_0^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim \|f\|_{\mathcal{K}_1(\mathbb{P}_0^d)} n^{-\frac{1}{2} - \frac{1}{2d}}. \quad (27)$$

We get rate $O(n^{-\frac{1}{2} - \frac{1}{d}})$ for

- ReLU and ReLU² (Klusowski & Barron)
- all ReLU^k (Xu)

What are the optimal rates for ReLU^k dictionaries?

Ref: Makovoz (1998), Xu (2020), Klusowski & Barron (2018)

Optimal Rates

Theorem (Siegel, Xu)

For the ReLU^k dictionary \mathbb{P}_k^d , we get

$$\inf_{f_n \in \Sigma_n(\mathbb{P}_k^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim \|f\|_{\mathcal{K}_1(\mathbb{P}_k^d)} n^{-\frac{1}{2} - \frac{2k+1}{2d}}. \quad (28)$$

- In fact, $f_n \in \Sigma_n^M(\mathbb{P}_k^d)$, with $M \lesssim \|f\|_{\mathcal{K}_1(\mathbb{P}_k^d)}$
- Rate is optimal (up to log factors) for *stable* approximation
- Holds more generally for any smoothly parameterizable dictionary \mathbb{D}
- Rate has been obtained in L^∞ for $k = 1$ (Matousek (1995), Bach (2017) and for $k = 0$ (Ma, Siegel, X 2022))

Proof uses piecewise polynomial approximation of the dictionary \mathbb{D}

Ref: Siegel & X (2022), Ma, Siegel, X (2022), Matousek (1995), Bach (2017)

Smoothly Parameterized Dictionaries

- Let $U \subset \mathbb{R}^d$ be an open set and $f : U \rightarrow \mathbb{R}$. Let $s = k + \alpha$ ($k \geq 0, \alpha \in (0, 1]$). Recall

$$|f|_{Lip(s, L^\infty(U))} := \sup_{x \neq y \in U} \frac{|D^k f(x) - D^k f(y)|}{|x - y|^\alpha}. \quad (29)$$

- Now consider a map $\mathcal{P} : U \rightarrow X$.

Definition

The map \mathcal{P} is of smoothness class s if for any $\xi \in X^*$ we have, letting $f_\xi(x) = \langle \mathcal{P}(x), \xi \rangle$,

$$|f_\xi|_{Lip(s, L^\infty(U))} \leq C \|\xi\|_{X^*}. \quad (30)$$

- Extended to smooth manifolds via charts

- Ref: Siegel & Xu 2022

Examples of Smoothly Parameterized Dictionaries

- Consider the Heaviside activation function

$$\sigma_0(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0, \end{cases}$$

and the map $\mathcal{P}_0^d : S^{d-1} \times [\alpha, \beta] \rightarrow L^p(\Omega)$ given by

$$\mathcal{P}_0^d(\omega, b) = \sigma_0(\omega \cdot x + b). \quad (31)$$

- Claim: \mathcal{P}_0^d is of smoothness class $\frac{1}{p}$. Indeed,

$$\|\sigma_0(\omega \cdot x + b) - \sigma_0(\omega' \cdot x + b')\|_{L^p(B_1^d)}^p \lesssim |\omega - \omega'| + |b - b'| \quad (32)$$

Examples of Smoothly Parameterized Dictionaries

- Consider the ReLU^k activation function

$$\sigma_1(x) = \begin{cases} x^k & x > 0 \\ 0 & x \leq 0, \end{cases}$$

and the map $\mathcal{P}_1^d : S^{d-1} \times [\alpha, \beta] \rightarrow L^p(\Omega)$ given by

$$\mathcal{P}_k^d(\omega, b) = \sigma_k(\omega \cdot x + b). \quad (33)$$

- Taking k derivatives, we get back to σ_0
- This implies that \mathcal{P}_k^d is of smoothness class $k + \frac{1}{p}$.

Main Theorem Upper Bounds

Theorem (Siegel & X 2022)

Let X be a type-2 Banach space. Suppose that \mathbb{D} is a parameterized by a smooth compact d -dimensional manifold \mathcal{M} with smoothness order s . Then for $f \in B_1(\mathbb{D})$ we have

$$\inf_{f_n \in \Sigma_n(\mathbb{D})} \|f - f_n\|_X \lesssim n^{-\frac{1}{2} - \frac{s}{d}}, \quad (34)$$

where the implied constant is independent of n .

- For ReLU^k networks, i.e. $\mathbb{D} = \mathbb{P}_k^d$, we get the rate $n^{-\frac{1}{2} - \frac{2k+1}{2d}}$ in $L^2(\Omega)$.
- Previous best rate was $n^{-\frac{1}{2} - \frac{1}{d}}$ in $L^2(\Omega)$ when $k > 1$.

Sketch of Proof

- Step 1: Reduce to the case where $\mathcal{M} = [0, 1]^d$.
- Step 2: Subdivide the cube into n subcubes C_1, \dots, C_n with diameter $O(n^{-\frac{1}{d}})$.
- Step 3: Form a piecewise polynomial interpolation of the parameterization \mathcal{P} on each of the cubes C_i using polynomials of degree k . On C_i , this interpolation has the form

$$P_k(z) = \sum_{l=1}^P \mathcal{P}(c_l) p_l^k(z). \quad (35)$$

- Step 4: Decompose $f = \sum_{i=1}^N a_i \mathcal{P}(z_i)$ as

$$f = \sum_{i=1}^N a_i P_k(z_i) + \sum_{i=1}^N a_i (\mathcal{P}(z_i) - P_k(z_i)). \quad (36)$$

Sketch of Proof (cont.)

- Step 5: Note that regardless of N , we have

$$\sum_{i=1}^N a_i P_k(z_i) \in \Sigma_{Pn}(\mathbb{D})! \quad (37)$$

- Step 6: Use a Bramble-Hilbert type lemma to prove the remainder bound (here we use smoothness of the parameterization)

$$\|\mathcal{P}(z) - P_k(z)\|_X \lesssim n^{-\frac{s}{d}}. \quad (38)$$

- Finally, apply original sampling argument to

$$\sum_{i=1}^N a_i (\mathcal{P}(z_i) - P_k(z_i)) \quad (39)$$

to complete the proof.

'Algorithmically' Achieving the Rate

How can we construct optimal shallow networks?

- Orthogonal Greedy Algorithm

$$f_0 = 0, g_k = \arg \max_{g \in \mathbb{D}} \langle r_{k-1}, g \rangle, f_k = P_k f \quad (40)$$

- ▶ $r_k = f - f_k$ is the residual
- ▶ P_k denotes the orthogonal projection onto the space spanned by g_1, \dots, g_k
- For general dictionaries \mathbb{D} , get $O(n^{-\frac{1}{2}})$ convergence
 - ▶ Not optimal for ReLU^k!

Can this be improved?

Ref: DeVore & Temlyakov (1996)

Optimal Orthogonal Greedy Convergence Rates

Theorem (Siegel & X 2022)

Let the iterates f_n be given by the orthogonal greedy algorithm, where $f \in \mathcal{K}_1(\mathbb{P}_k^d)$. Then we have

$$\|f_n - f\|_{L^2} \lesssim \|f\|_{\mathcal{K}_1(\mathbb{P}_k^d)} n^{-\frac{1}{2} - \frac{2k+1}{2d}}. \quad (41)$$

- Implies that the OGA trains optimal neural networks
- Downside: no stability, i.e. $\|f_n\|_{\mathcal{K}_1(\mathbb{D})}$ may be arbitrarily large!

Should we be excited?

- 1 NN has **SUPER**-approximation property!
- 2 NN **breaks** curse-of-dimensionality?

Caution:

We should not get too excited by such a “dimension-independent” result!

Example: a network of 3 parameters

$$\Sigma_3^{\text{coscos}} = \left\{ C \cos(t \cos(\lfloor Kx \rfloor)), C, t, K \in \mathbb{R} \right\}, \quad (42)$$

$$\lfloor x \rfloor = \text{largest integer that is } \leq x. \quad (43)$$

Theorem

For any continuous function g on $[0, 1]$ and any $\epsilon > 0$, there exist $C, t, K \in \mathbb{R}$ such that

$$\|g - f(\circ; C, t, K)\|_{L^\infty([0,1])} < \epsilon. \quad (44)$$

- This theorem means

$$\inf_{u_3 \in \Sigma_3^{\text{coscos}}} \|u - u_3\| = 0 = \mathcal{O}(3^{-\infty}). \quad (45)$$

- Three parameters suffice to capture any function!
- Parameters must be extremely large to obtain high accuracy
 - ▶ Number of parameters is not a priori useful notion
 - ▶ Cannot be specified with a fixed number of bits
 - ▶ Not *encodable*!
- Shen, Z., Yang, H. & Zhang, S. (2021)

Proof

- 1 Choose $C = \|g\|_{L^\infty([0,1])}$. We assume next that $\|g\|_{L^\infty([0,1])} \leq 1$.
- 2 Choose $K \in \mathbb{N}$ sufficiently large such that

$$\max_{x \in \left[\frac{j}{K}, \frac{j+1}{K}\right]} \left| g(x) - g\left(\frac{j}{K}\right) \right| < \frac{\epsilon}{2}, \quad j = 0, 1, \dots, K.$$

- 3 The set $\{\cos 0, \cos 1, \dots, \cos(K)\}$ is linearly independent over \mathbb{Q} since $\cos 1$ is transcendental.
- 4 $\{t(\cos 0, \dots, \cos(K)) : t \in \mathbb{R}\}$ is dense in $\mathbb{R}^{K+1}/(2\pi\mathbb{Z})^{K+1}$. Namely there exists some $t \in \mathbb{R}$ and $\mathbf{m} \in \mathbb{Z}^{K+1}$ such that

$$\|(t \cos 0, \dots, t \cos(K)) + 2\pi \mathbf{m} - \mathbf{y}\|_{L^\infty([0,2\pi]^{K+1})} < \frac{\epsilon}{2}.$$

for $\mathbf{y} = \left(\arccos\left(g\left(\frac{0}{K}\right)\right), \dots, \arccos\left(g\left(\frac{K}{K}\right)\right)\right)$.

Now for any $x \in [0, 1]$, there exists some $0 \leq j \leq K$ such that $x \in \left[\frac{j}{K}, \frac{j+1}{K}\right)$. Thus

$$\begin{aligned} |f(x; 1, t, K) - g(x)| &= \left| f(x; 1, t, K) - g\left(\frac{j}{K}\right) \right| + \left| g\left(\frac{j}{K}\right) - g(x) \right| \\ &\leq \left| \cos(t \cos(j)) - g\left(\frac{j}{K}\right) \right| = \left| \cos(t \cos(j) + 2\pi m_j) - \cos\left(\arccos\left(g\left(\frac{j}{K}\right)\right)\right) \right| + \frac{\epsilon}{2} \\ &\leq \left| t \cos(j) + 2\pi m_j - \arccos\left(g\left(\frac{j}{K}\right)\right) \right| + \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

This is

$$\|f(\cdot; 1, t, K) - g\|_{L^\infty([0,1])} < \epsilon.$$

- 1 Shallow neural networks
- 2 Dictionary and variation spaces
- 3 Approximation properties of shallow neural networks
- 4 Metric Entropy**
- 5 Summary

Encodability: metric entropy

Definition (Kolmogorov)

Let X be a Banach space and $B \subset X$. The metric entropy numbers of B , $\epsilon_n(B)_X$ are given by

$$\epsilon_n(B)_X = \inf\{\epsilon : B \text{ is covered by } 2^n \text{ balls of radius } \epsilon\}. \quad (46)$$

- For example, the interval $[0, 1]$ can be covered by 2^n balls of radius $\frac{1}{2^{n+1}}$. But it cannot be covered by 2^n balls of radius less than this. So $\epsilon_n([0, 1]) = \frac{1}{2^{n+1}}$. For the d -dimensional cube $[0, 1]^d$, the metric entropy (with respect to the ℓ^∞ norm) is $\epsilon_n([0, 1]^d) \simeq \frac{1}{2^{n/d}}$.
- $\epsilon_n(B)_K$ measures how accurately elements of B can be specified with n bits, i.e. $\epsilon_n(B)_K$ measures best approximation by \mathcal{F}_n which is encodable with n bits
- High-dimensional balls do not always have larger entropy than low-dimensional balls: For $B \in \mathbb{R}^d$ is the unit ball $\epsilon_n(rB)_X = r\epsilon_n(B)_X$. The entropy of rB can be small when r is small.
- Gives fundamental limit for any (digital) numerical algorithm
- Gives fundamental limit on stable (i.e. Lipschitz) approximation methods
- Curse of dimensionality: for unit ball B_p^s in Sobolev space $W^{s,p}(\Omega) : \epsilon_n(B_p^s)_{L^p(\Omega)} \sim n^{-\frac{s}{d}}$
- In high dimensions, we need novel function classes with small metric entropy!

Ref: Birman & Solomyak (1967), Mhaskar. H. N., Narcowich, F. J, and Ward. J. D. (2004), Cohen, Devore, Petrova, Wojtaszczyk (2021)

No curse of dimensionality: polynomial & kernel

Theorem

$$\inf_{u_n \in P_n} \|u - u_n\| \approx n^{-\frac{s}{d}} \|u\|_{H^s(\Omega)}, \quad (47)$$

where $\Omega = [0, 1]^d$, $u \in H^s(\Omega)$, P_n is the space of polynomials on Ω with n degree of freedom.

Theorem

Let Q be a Gaussian kernel and $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ be appropriately distributed, for any $s > \frac{d}{2}$ we have

$$\inf_{u_n \in Q_n} \|u - u_n\| \lesssim n^{-\frac{s}{d}} \|u\|_{H^s}, \text{ where } Q_n = \text{span}\{Q(x, x_i)\}_{i=1}^n \quad (48)$$

No curse of dimensionality in both cases for sufficiently smooth functions:

$$\inf_{u_n} \|u - u_n\| \lesssim n^{-\frac{1}{2}} \|u\|_{H^{d/2}}. \quad (49)$$

- DeVore, R. A., & Lorentz, G. G. (1993), Mhaskar. H (1995), Arcangéli, R., López de Silanes, M. C., & Torrens, J. J. (2007), Narcowich. F. J, Ward. J. D., and Wendland. H (2006); Batlle, P., Chen, Y., Hosseini, B., Owhadi, H., & Stuart, A. M. (2023).

Entropy for classical spaces

Unit ball in Sobolev spaces

Theorem (Birman-Solomyak, 1967)

Let $\Omega = [0, 1]^d$. For $1 \leq p, q \leq \infty$ and $s/d > 1/q - 1/p$, the entropy of the unit ball in the Sobolev space $W^s(L_q([0, 1]^d))$ is estimated as

$$\epsilon_n(B_q^s)_{L^p(\Omega)} \approx n^{-\frac{s}{d}} \quad (50)$$

Analytic functions

Theorem (Komogorov, 1958)

Let $\mathcal{A}^d(K, G)$ consists of functions analytic in a domain (connected open bounded set) $G \subset \mathbb{C}^d$ with $|f(z)| \leq 1$ in G . Let K be a compact subset of G with nonempty interior. Then

$$\log \left(1/\epsilon_n(\mathcal{A}^d)_{L^\infty(K)} \right) \approx n^{\frac{1}{d+1}}. \quad (51)$$

Metric Entropy of Dictionary Spaces

What are the metric entropies of $\mathcal{K}_1(\mathbb{P}_k^d)$?

Theorem (Siegel & Xu 2022)

The metric entropies of \mathbb{P}_k^d and \mathbb{F}_s^d satisfy

$$\epsilon_n(B_1(\mathbb{P}_k^d)) \asymp n^{-\frac{1}{2} - \frac{2k+1}{2d}}, \quad \epsilon_n(B_1(\mathbb{F}_s^d)) \asymp n^{-\frac{1}{2} - \frac{s}{d}} \quad (52)$$

- No curse of dimensionality (in terms of metric entropy)!
- However, there appears to be an *algorithmic* curse of dimensionality
 - ▶ We have not found an efficient way to search over the dictionary \mathbb{P}_k^d

Ref: Siegel & X (2022)

- 1 Shallow neural networks
- 2 Dictionary and variation spaces
- 3 Approximation properties of shallow neural networks
- 4 Metric Entropy
- 5 Summary

Summary

- Shallow neural networks and its basic approximation properties
- Dictionary and variation spaces
- Approximation theory for shallow neural networks
- Metric entropy