Approximation and non-parametric estimation of functions over high-dimensional spheres via deep ReLU networks

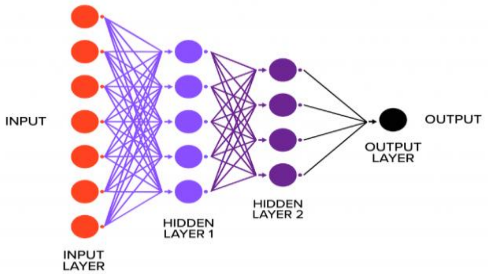
Namjoon Suh, Tian-Yi Zhou, Xiaoming Huo

Georgia Tech, ISyE

0. Two objects : Approximation error and excess risk

Approximation Error of neural network : how well neural networks approximates the functions in certain class.

$$\| \tilde{f} - f_\rho \|_\infty := inf_{\tilde{f} \in \mathcal{F}} sup_{f_\rho \in \mathrm{W}^{\mathrm{r}}_\infty(X)} |\tilde{f}(x) - f_\rho(x)|$$



The error is characterized by...

1. The complexity of neural network class $\mathcal{F}$
   - Three components : **depth (L)**, **width (W), number of units (U)** (Bartlett & Anthony, 1999).
   - In my work, $\mathcal{F}$ is set as fully-connected neural network with **sparse** weight.

2. The specific function space where the ground-truth belongs. (ex: **Holder**, Besov, **Sobolev,** RKHS, etc.)

$\mathrm{W}^{\mathrm{r}}_\infty(X)$
Holder space

1. $X = C^d$ (d-dimensional cube)

$\underbrace{\text{Mhaskar (1996), Yarotsky (2017), Peterson \& Voigtlaender (2018), Schmidt-Hieber (2020)}}_{\text{FNN}}$

2. $X = S^{d-1} := \{x \in \mathbb{R}^d : \| x \|_2 = 1\}$

Fang et al (2020), Feng et al (2021) : CNN

Above literature studies fixed ε-approximation accuracy, and express **L, W,** and **U** w.r.t. $\varepsilon \in (0,1)$ for fixed d.

My work : under the setting $f_\rho \in \mathrm{W}^{\mathrm{r}}_\infty(S^{d-1})$, letting $d \to \infty$, the approximation error is expressed in d.

0. Two objects : Approximation error and excess risk

Excess Risk of estimator : how well neural networks estimates the underlying functions with noisy observations $\{x_i, y_i\}_{i=1}^{n}$, where they are generated from...

$$y_i = f_\rho(x_i) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

It is assumed $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}(S^{d-1})$, and it is estimated through neural network, which is a minimizer of...

$$\hat{f}_n = \underset{f : f \in \mathcal{F}}{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \right\}$$

"Regularized estimator"

Further clip the estimator via the projection operator (Suzuki, 2018, Fang & Cheng, 2022, Oono & Suzuki, 2019) :

$$\pi_B f(x) \coloneqq \begin{cases} f(x), & \text{if } |f(x)| \leq B \\ B, & \text{if } f(x) > B \\ -B, & \text{if } f(x) < -B \end{cases}$$

We are interested in bounding **excess risk** defined as :

$$\mathcal{E}(\pi_B \hat{f}_n) - \mathcal{E}(f_\rho) = E_{(X,y)\sim\rho}\left[(\mathrm{y} - \pi_B \hat{f}_n(\mathrm{x}))^2\right] - E_{(X,y)\sim\rho}\left[(\mathrm{y} - f_\rho(\mathrm{x}))^2\right] = E_{X\sim\rho_X}\left[(\pi_B \hat{f}_n(\mathrm{x}) - f_\rho(\mathrm{x}))^2\right]$$

My work : under the setting $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}(S^{d-1})$, letting $d \to \infty$, the bound on excess risk is expressed in n and d.

1. Deep ReLU networks and Sobolev Space on Sphere

A deep ReLU network with a "***depth***" L and a "***width vector***" $\boldsymbol{p} = (\boldsymbol{p_0}, \boldsymbol{p_1}, \dots, \boldsymbol{p_{L+1}}) \in \mathbb{R}^{\boldsymbol{L+2}}$ is defined as :

$$\boxed{\tilde{f} : S^{d-1} \to \mathbb{R}, \qquad x \to \tilde{f}(x) = W_L \sigma_{V_L} W_{L-1} \sigma_{V_{L-1}} \dots \sigma_{V_1} W_1 x} \quad \bigstar$$

where $W_i \in \mathbb{R}^{P_{i+1} X P_i}$ is weight matrix and $v_i \in \mathbb{R}^{P_i}$ is a shift vector on ReLU activation $\sigma_{v_i}(x) = \max(x - v_i, 0)$.

The neural network class $\mathcal{F}$ we consider is written as :

$$\boxed{\mathcal{F}(L, \boldsymbol{p}, \mathcal{N}) := \left\{\tilde{f} \text{ of the form } \bigstar : \textstyle\sum_{j=1}^{L} \| W_j \|_0 + \| v_j \|_0 \le \mathcal{N}\right\}}$$

where $\| W_j \|_0$ and $\| v_j \|_0$ denotes the number of non-zero entries in W_j and v_j.

For $\mathrm{x} \in \mathrm{S}^{\mathrm{d}-1}$, $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}(\mathrm{S}^{\mathrm{d}-1}) \subseteq \mathcal{L}_2(\mathrm{S}^{\mathrm{d}-1})$, it can be written as :

$$\boxed{f_\rho(x) = \sum_{k=0}^{\infty} Proj_k(f_\rho)(x) = \sum_{k=0}^{\infty} \sum_{l=1}^{\mathcal{N}(k,d)} \hat{f}_{k,l} Y_{k,l}(x)}$$

Dimension of $\mathcal{H}_k^d$

Orthonormal basis of $\mathcal{H}_k^d$ of degree $l \le k$.

Fourier Coefficient of f^* given by $< f_\rho, Y_{k,l}(x) >_{\mathcal{L}_2(\mathrm{S}^{\mathrm{d}-1})}$.

For $\mathrm{x} \in \mathrm{S}^{\mathrm{d}-1}$, $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}(\mathrm{S}^{\mathrm{d}-1}) \subseteq \mathcal{L}_2(\mathrm{S}^{\mathrm{d}-1})$ is defined as :

$$\boxed{\| f_\rho \|_{\mathrm{W}_\infty^{\mathrm{r}}(\mathrm{S}^{\mathrm{d}-1})} = \| \underbrace{\left(-\Delta_{\mathrm{S}^{\mathrm{d}-1}} + I\right)}_{}^{r/2} f_\rho \|_p < \infty}$$

Laplace-Beltrami Operator : (Hessian Operator on Euclidean Space).

1. Deep ReLU networks and Sobolev Space on Sphere

A deep ReLU network with a "*depth*" L and a "*width vector*" $\boldsymbol{p} = (p_0, p_1, \ldots, p_{L+1}) \in \mathbb{R}^{L+2}$ is defined as :

$$\boxed{\tilde{f} : S^{d-1} \to \mathbb{R}, \qquad x \to \tilde{f}(x) = W_L \sigma_{V_L} W_{L-1} \sigma_{V_{L-1}} \ldots \sigma_{V_1} W_1 x} \quad ✡$$

where $W_i \in \mathbb{R}^{P_{i+1} X P_i}$ is weight matrix and $v_i \in \mathbb{R}^{P_i}$ is a shift vector on ReLU activation $\sigma_{v_i}(x) = \max(x - v_i, 0)$.

The neural network class $\mathcal{F}$ we consider is written as :

$$\boxed{\mathcal{F}(L, \boldsymbol{p}, \mathcal{N}) \coloneqq \left\{\tilde{f} \text{ of the form } ✡ : \textstyle\sum_{j=1}^{L} \| W_j \|_0 + \| v_j \|_0 \le \mathcal{N}\right\}}$$

where $\| W_j \|_0$ and $\| v_j \|_0$ denotes the number of non-zero entries in W_j and v_j.

For $x \in S^{d-1}$, $f_\rho \in W^r_\infty(S^{d-1}) \subseteq \mathcal{L}_2(S^{d-1})$, it can be written as :

$Proj_k(f_\rho) \in \mathcal{H}^d_k$

Dimension of $\mathcal{H}^d_k$

$$\boxed{f_\rho(x) = \sum_{k=0}^{\infty} Proj_k(f_\rho)(x) = \sum_{k=0}^{\infty} \sum_{l=1}^{\mathcal{N}(k,d)} \hat{f}_{k,l} Y_{k,l}(x)}$$

Orthonormal basis of $\mathcal{H}^d_k$ of degree $l \le k$.

Fourier Coefficient of f^* given by $< f_\rho, Y_{k,l}(x) >_{\mathcal{L}_2(S^{d-1})}$.

For $x \in S^{d-1}$, $f_\rho \in W^r_\infty(S^{d-1}) \subseteq \mathcal{L}_2(S^{d-1})$ is defined as :

$$\boxed{\| f_\rho \|_{W^r_\infty(S^{d-1})} = \| \underbrace{\left(-\Delta_{S^{d-1}} + I\right)}_{\text{Laplace-Beltrami Operator : (Hessian Operator on Euclidean Space).}}{}^{r/2} f_\rho \|_\infty < \infty}$$

2. Approximation result & Comparisons with existing literature

Our Result : Let $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}(S^{d-1})$. For $0 < \alpha, \beta, \gamma < 1$, and some constants $C, C' > 0$ independent with d:

1. When $\boldsymbol{r = O(d)}$ as $d \to \infty$, then there exists a neural network $\tilde{f} \in \mathcal{F}(L, \boldsymbol{p}, \mathcal{N})$ with depth $L = O(d^\gamma log_2 d)$, $W = O(d^\alpha)$, $\mathcal{N} = O(d^{\max\{\alpha+\gamma,1\}})$, with approximation rate :

$$\boxed{\| \tilde{f} - f_\rho \|_\infty \leq C \| f_\rho \|_{\mathrm{W}_\infty^{\mathrm{r}}(\mathrm{S}^{\mathrm{d}-1})} d^{-d^\beta}}$$

2. When $\boldsymbol{r = O(1)}$ as $d \to \infty$, then there exists a neural network $\tilde{f} \in \mathcal{F}(L, \boldsymbol{p}, \mathcal{N})$ with depth $L = O(d^\gamma log_2 d)$, $W = O((9d)^d)$, $\mathcal{N} = O((9d)^d)$, with approximation rate :

$$\boxed{\| \tilde{f} - f_\rho \|_\infty \leq C' \| f_\rho \|_{\mathrm{W}_\infty^{\mathrm{r}}(\mathrm{S}^{\mathrm{d}-1})} d^{-\alpha r}}$$

Implication 1 : When the given function smoothness increases from $\boldsymbol{r = O(1)}$ to $\boldsymbol{r = O(d)}$, the required width for the approximation becomes narrower, while the smoothness has little effect on the depth of network.

Implication 2 : When $\boldsymbol{r = O(d)}$, the deep ReLU FNN avoids the ***"Curse of dimensionality"***, requiring at most $\mathcal{N} = O(d^2)$, with a very sharp approximation error rate.

Implication 3 : The same observation is not found in approximation theory result in $f_\rho \in \mathrm{W}_\infty^{\mathrm{r}}([0,1]^d)$.

2. Approximation result & Comparisons with existing literature

Our Result : Let $f_\rho \in W^r_\infty(S^{d-1})$. For $0 < \alpha, \beta, \gamma < 1$, and some constants $C, C' > 0$ independent with d:

1. When $\boldsymbol{r = O(d)}$ as $d \to \infty$, then there exists a neural network $\tilde{f} \in \mathcal{F}(L, \boldsymbol{p}, \mathcal{N})$ with depth $L = O(d^\gamma log_2 d)$, $W = O(d^\alpha)$, $\mathcal{N} = O(d^{\max\{\alpha+\gamma,1\}})$, with approximation rate :

$$\boxed{\| \tilde{f} - f_\rho \|_\infty \le C \| f_\rho \|_{W^r_\infty(S^{d-1})} \, d^{-d^\beta}}$$

2. When $\boldsymbol{r = O(1)}$ as $d \to \infty$, then there exists a neural network $\tilde{f} \in \mathcal{F}(L, \boldsymbol{p}, \mathcal{N})$ with depth $L = O(d^\gamma log_2 d)$, $W = O((9d)^d)$, $\mathcal{N} = O((9d)^d)$, with approximation rate :

$$\boxed{\| \tilde{f} - f_\rho \|_\infty \le C' \| f_\rho \|_{W^r_\infty(S^{d-1})} \, d^{-\alpha r}}$$

Theorem (Schimdt-Hieber, 2020)

For any function $f \in W^r_\infty([0,1]^d)$ and let $K > 0$ be the radius of Hölder ball. Then, for any integers $m \ge 1$ and $N^H \ge (r+1)^d \vee (K+1)e^d$, there exists a network

$$\tilde{f}^H \in \mathcal{F}^H(L, (d, 6(d+\lceil r \rceil)N^H, \ldots, 6(d+\lceil r \rceil)N^H, 1), \mathcal{N}^H)$$

with depth $L = 8 + (m+5)(1 + \lceil \log_2(d \vee r) \rceil)$ and the number of parameters $\mathcal{N}^H \le 141(1+d+r)^{3+d} N^H (m+6)$, such that

$$\left\| f - \tilde{f}^H \right\|_\infty \le (2K+1)(1+d^2+r^2)6^d (N^H) 2^{-m} + K 3^r (N^H)^{-\frac{r}{d}}.$$

1. As either r or d increases, the width W and number of active parameters $\mathcal{N}$ increases.

2. The approximation error : can't observe the interactions between r and d, specifically in the first term.

3. Bounds on Excess risk

For some constants $C > 0$, we have a bound for the excess risk: $\boldsymbol{\mathcal{E}(\pi_B \hat{f}_n) - \mathcal{E}(f_\rho)}$!!

	Theorem 3.		Theorem 4.
Function class	$W_\infty^r(\mathcal{S}^{d-1})$		$W_\infty^r([0,1]^d)$
Smoothness r	$\mathcal{O}(d)$	$\mathcal{O}(1)$	$\forall r > 0$
Upper-bound on $\mathcal{N}$	$\mathcal{O}(nd)$	$\mathcal{O}(nd)$	$\tilde{\mathcal{O}}((d+r)^d)$
Estimation error rate	$\tilde{\mathcal{O}}(d^C \cdot n^{-\frac{4r}{4r+3d}})$	$\tilde{\mathcal{O}}((\frac{6}{\pi e})^{\frac{d}{2}} d^d \cdot n^{-\frac{4r}{4r+3d}})$	$\tilde{\mathcal{O}}((d+r)^d \cdot n^{-\frac{2r}{2r+d}})$

1. Bounds on excess risks are written in terms of n and d :
   - Note the bounds for the estimation error of $f_\rho \in W_\infty^r(S^{d-1})$ is sub-optimal, whereas the error bound for $f_\rho \in W_\infty^r([0,1]^d)$ is optimal. (Donoho & Johnstone, 1998)

   - Note that in case of $f_\rho \in W_\infty^r(S^{d-1})$, the order of d in the constant factors are dependent on r, whereas $f_\rho \in W_\infty^r([0,1]^d)$ has no such dependency.

2. Interestingly, in case of $f_\rho \in W_\infty^r(S^{d-1})$, only $O(d^2)$ for $\mathcal{N}$ are required for all $r > 0$, whereas, in case of $f_\rho \in W_\infty^r([0,1]^d)$, the upper-bound for $\mathcal{N}$ is exponentially dependent on d for all $r > 0$.