
Approximation Properties of Neural Network Function Class

Given an activation function

$$(20.1) \quad \sigma : \mathbb{R}^1 \mapsto \mathbb{R}^1$$

We consider the following shallow neural network function class

$$(20.2) \quad \Sigma_d(\sigma) = \text{span} \left\{ \sigma(\omega \cdot x + \theta) : \omega \in \mathbb{R}^d, \theta \in \mathbb{R} \right\}.$$

In this chapter, we will two theorems as follows.

Theorem 123. *Let σ be a non-polynomial Riemann integrable function and $\sigma \in L_{loc}^\infty(\mathbb{R})$. Then $\Sigma_d(\sigma)$ is dense in $C(\mathbb{R}^d)$.*

Theorem 124. *Let $\Omega \subset \mathbb{R}^d$ be a bounded set, $\sigma \in W^{m,\infty}(\mathbb{R})$ that has compact support such that*

$$(20.3) \quad \hat{\sigma}(a) \neq 0, \text{ for some } a \neq 0.$$

Then

$$(20.4) \quad \inf_{f_n \in \Sigma_n} \|f - f_n\|_{H^m(\Omega)} \leq C(d, m) n^{-\frac{1}{2}} \int_{\mathbb{R}^d} (1 + |\omega|)^{m+1} |\hat{f}(\omega)|$$

20.1 Non-polynomial as activation function

Lemma 127. *Let $\sigma \in C^\infty(\mathbb{R})$ and assume σ is not a polynomial. Then $\Sigma_n(\sigma)$ is dense in $C(\mathbb{R}^n)$.*

Proof. Since $\sigma \in C^\infty(\mathbb{R})$, and $[\sigma((\omega + h e_j) \cdot x + \theta) - \sigma(\omega \cdot x + \theta)]/h \in \Sigma_n(\sigma)$ for every ω, θ and $h \neq 0$, it follows that $\frac{\partial}{\partial \omega_j} \sigma(\omega \cdot x + \theta) \in \overline{\Sigma}_n(\sigma)$ for all $j = 1 : n$. By the same argument $\frac{\partial^k}{\partial \omega_j^k} \sigma(\omega \cdot x + \theta) \in \overline{\Sigma}_n(\sigma)$ for all $k \in \mathbb{N}$, $j = 1 : n$, $\omega \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$.

Now $\frac{\partial^k}{\partial \omega_j^k} \sigma(\omega \cdot x + \theta) = x_j^k \sigma^{(k)}(\omega \cdot x + \theta)$, and since σ is not a polynomial there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. Take $\omega = 0$ and $\theta = \theta_k$, we then have $x_j^k \in \overline{\Sigma}_n(\sigma)$. Similarly, for all polynomials of the form $x_1^{k_1} \cdots x_n^{k_n}$, we can get them by taking the corresponding partial derivatives.

This implies that $\overline{\Sigma}_n(\sigma)$ contains all polynomials. By Weierstrass's Theorem it follows that $\overline{\Sigma}_n(\sigma)$ contains $C(K)$ for each compact $K \subset \mathbb{R}^n$. That is $\Sigma_n(\sigma)$ is dense in $C(\mathbb{R}^n)$. \square

Theorem 125. *Let σ be a non-polynomial Riemann integrable function and $\sigma \in L_{loc}^\infty(\mathbb{R})$. Then $\Sigma_1(\sigma)$ is dense in $C(\mathbb{R})$.*

Proof. Consider the mollifier η

$$\eta(x) = \begin{cases} C \exp\left(\frac{1}{|x|^2 - 1}\right), & |x| < 1, \\ 0, & |x| \geq 1. \end{cases}$$

here C is selected so that $\int_{\mathbb{R}} \eta dx = 1$.

Set $\eta_\epsilon = \frac{1}{\epsilon} \eta\left(\frac{x}{\epsilon}\right)$. Then consider σ_{η_ϵ}

$$(20.5) \quad \sigma_{\eta_\epsilon}(x) := \sigma * \eta_\epsilon(x) = \int_{\mathbb{R}} \sigma(x - y) \eta_\epsilon(y) dy$$

It can be seen that $\sigma_{\eta_\epsilon} \in C^\infty(\mathbb{R})$. Following the proof in the previous proposition, we want to show that $\overline{\Sigma}_1(\sigma)$ contains all polynomials.

The first step is to show that $\overline{\Sigma}_1(\sigma_{\eta_\epsilon}) \subset \overline{\Sigma}_1(\sigma)$, which can be done easily by checking the Riemann sum of $\sigma_{\eta_\epsilon}(x) = \int_{\mathbb{R}} \sigma(x - y) \eta_\epsilon(y) dy$ is in $\overline{\Sigma}_1(\sigma)$.

Then it suffices to show that there exists θ_k and σ_{η_ϵ} such that $\sigma_{\eta_\epsilon}^{(k)}(\theta_k) \neq 0$ for each k. If not, then there must be k_0 such that $\sigma_{\eta_\epsilon}^{(k_0)}(\theta) = 0$ for all $\theta \in \mathbb{R}$ and all $\epsilon > 0$. Thus σ_{η_ϵ} 's are all polynomials with degree at most $k_0 - 1$. In particular, It is known that $\eta_\epsilon \in C_0^\infty(\mathbb{R})$ and $\sigma * \eta_\epsilon$ uniformly converges to σ on compact sets in \mathbb{R} and $\sigma * \eta_\epsilon$'s are all polynomials of degree at most $k_0 - 1$. Polynomials of a fixed degree form a closed linear subspace, therefore σ is also a polynomial of degree at most $k_0 - 1$, which leads to contradiction. \square

20.2 Convergence rates in Sobolev norms

20.2.1 Compactly supported activation function

Given a bounded domain $\Omega \subset \mathbb{R}^d$, we consider the function

$$f : \Omega \mapsto \mathbb{R}$$

Let

$$f^e : \mathbb{R}^d \mapsto \mathbb{R}$$

be any extension of f so that

$$f^e|_{\Omega} = f(x), \quad x \in \Omega.$$

Most time, we will drop the superscript “ e ” to still use f to denote an extension of f .

Consider the Fourier transform:

$$(20.6) \quad \hat{f}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega \cdot x} f(x) dx \quad \forall \omega \in \mathbb{R}^d.$$

Let $x_{\Omega} \in \Omega$ is such that

$$(20.7) \quad x_{\Omega} \in \arg \min_{y \in \Omega} (\max_{x \in \Omega} |x - y|)$$

We may call

$$(20.8) \quad r_{\Omega} = \max_{x \in \Omega} |x - x_{\Omega}|$$

the radius of Ω .

Using the Fourier inversion formula, we can have a Fourier representation of $f(x)$ as follows

$$(20.9) \quad f(x) = \int_{\mathbb{R}^d} e^{i\omega \cdot x} \hat{f}(\omega) d\omega \quad \forall x \in \Omega,$$

Let us write

$$(20.10) \quad \hat{f}(\omega) = e^{i\beta_1(\omega)} |\hat{f}(\omega)|.$$

Let σ be the activation function with compact support and bounded derivatives up to order m , that is

$$(20.11) \quad \|\sigma\|_{m,\infty} := \max_{0 \leq \alpha \leq m} \sup_{x \in \mathbb{R}} |\sigma^{(\alpha)}(x)| \leq \infty$$

Set

$$(20.12) \quad \tilde{x} = x - x_{\Omega}.$$

Choose $a \neq 0$ such that $\hat{\sigma}(a) \neq 0$, by Fourier transform, we have:

$$\hat{\sigma}(a) = \frac{1}{2\pi} \int_{\mathbb{R}} \sigma(y) e^{-iay} dy = \frac{1}{2\pi} \int_{\mathbb{R}} \sigma(\omega \cdot \tilde{x} + b) e^{-ia(\omega \cdot \tilde{x} + b)} db$$

and we can also write

$$\hat{\sigma}(a) = e^{i\beta_2(a)} |\hat{\sigma}(a)|.$$

Thus

$$(20.13) \quad |\hat{\sigma}(a)| = \frac{1}{2\pi} \int_{\mathbb{R}} \sigma((\omega \cdot \tilde{x} + b)) e^{-ia(\omega \cdot \tilde{x} + b) - i\beta_2(a)} db.$$

Then we can write

$$(20.14) \quad \begin{aligned} f(x) &= \int_{\mathbb{R}^d} e^{i\omega \cdot x} e^{i\beta_1(\omega)} |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} |a|^d e^{ia(\omega \cdot x)} e^{i\beta_1(a\omega)} |\hat{f}(a\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \frac{|a|^d}{|\hat{\sigma}(a)|} |\hat{\sigma}(a)| e^{ia(\omega \cdot x) + i\beta_1(a\omega)} |\hat{f}(a\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} \frac{|a|^d}{2\pi |\hat{\sigma}(a)|} \sigma(\omega \cdot \tilde{x} + b) e^{-i\beta_2(a)} e^{ia\omega \cdot x_\Omega - iab} db e^{i\beta_1(a\omega)} |\hat{f}(a\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} \sigma(\omega \cdot x + b) \frac{|a|^d}{2\pi |\hat{\sigma}(a)|} e^{i\beta_3(a, \omega, b)} |\hat{f}(a\omega)| db d\omega \end{aligned}$$

where $\beta_3(a, \omega, b) = a\omega \cdot x_\Omega - ab - \beta_2(a) + \beta_1(a\omega)$.

Let

$$\theta = (\omega, b).$$

Since f is real, we have

$$f(x) = \operatorname{Re} f(x) = \int_{\mathbb{R}^{d+1}} \kappa(\theta) \sigma(\omega \cdot \tilde{x} + b) |a|^d |\hat{f}(a\omega)| d\theta.$$

where

$$(20.15) \quad \kappa(\theta) \equiv \kappa(\omega, b) = \frac{\cos \beta_3(a, \omega, b)}{2\pi |\hat{\sigma}(a)|},$$

Notice that σ is compactly supported, assume that $\operatorname{supp}(\sigma) \subset [-M_1, M_1]$. Then we have:

$$(20.16) \quad f(x) = \int_{\mathbb{R}^{d+1}} \kappa(\theta) \mathbf{1}_{D_M} \sigma(\omega \cdot \tilde{x} + b) |a|^d |\hat{f}(a\omega)| d\theta$$

where

$$(20.17) \quad M = \max(M_1, r_\Omega)$$

$$(20.18) \quad D_M = \{\theta = (\omega, b) : |b| \leq (1 + |\omega|)M\}$$

Define

$$(20.19) \quad \gamma(f) = \int_{\mathbb{R}^{d+1}} \mathbf{1}_{D_M} (1 + |\omega|)^m |a|^d |\hat{f}(a\omega)| d\theta$$

Now that

$$(20.20) \quad \begin{aligned} f(x) &= \int_{\mathbb{R}^{d+1}} \kappa(\theta) \mathbf{1}_{D_M} \sigma(\omega \cdot \tilde{x} + b) |a|^d |\hat{f}(a\omega)| d\theta \\ &= \int_{\mathbb{R}^{d+1}} \frac{\kappa(\theta) \gamma(f)}{(1 + |\omega|)^m} \sigma(\omega \cdot \tilde{x} + b) \frac{(1 + |\omega|)^m \mathbf{1}_{D_M} |a|^d |\hat{f}(a\omega)|}{\gamma(f)} d\theta \\ &:= \int_{\mathbb{R}^{d+1}} \frac{\kappa(\theta) \gamma(f)}{(1 + |\omega|)^m} \sigma(\omega \cdot \tilde{x} + b) d\lambda \end{aligned}$$

where

$$d\lambda = \frac{(1 + |\omega|)^m \mathbf{1}_{D_m} |a|^{d_1} |\hat{f}(a\omega)|}{\gamma(f)} d\theta, \quad \int_{\mathbb{R}^{d+1}} d\lambda = 1,$$

and we can write

$$(20.21) \quad \tilde{f}(x) \equiv \frac{f(x)}{\gamma(f)} = \mathbb{E}(g(\theta; x))$$

where

$$(20.22) \quad g(\theta; x) = \frac{\kappa(\theta)}{(1 + |\omega|)^m} \sigma(\omega \cdot \tilde{x} + b)$$

And

$$(20.23) \quad D^\alpha \tilde{f}(x) = \mathbb{E}(D^\alpha g).$$

Let $\theta_i = (\omega_i, b_i)_{i=1}^n$ be independently drawn from the same distribution λ and let

$$\bar{g}(\theta_1, \dots, \theta_n) = \frac{1}{n} \sum_{i=1}^n g(\theta_i, x).$$

Then

$$\mathbb{E} \left(\sum_{|\alpha| \leq m} (D^\alpha \tilde{f} - D^\alpha \bar{g})^2 \right) = \sum_{|\alpha| \leq m} \mathbb{E} (D^\alpha \tilde{f} - D^\alpha \bar{g})^2 = \sum_{|\alpha| \leq m} \mathbb{E} (\mathbb{E}(D^\alpha g) - D^\alpha \bar{g})^2 \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

Taking the L^2 norm (with a probability measure) on both side of the above inequality and using Fubini's theorem,

$$(20.24) \quad \mathbb{E} (\|\tilde{f}(\cdot) - g(\theta, \cdot)\|_{H^m(\Omega)}^2) \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

Thus there exists $\theta_i = (\omega_i^*, b_i^*)_{i=1}^n$ and β_i such that

$$\|\tilde{f} - \frac{1}{n} \sum_{i=1}^n g(\theta_i; \cdot)\|_m^2 \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

$$\|f - f_n\|_m^2 \leq \frac{\gamma(f)^2}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

where

$$f_n(x) = \frac{\gamma(f)}{n} \sum_{i=1}^n g(\theta_i; \cdot) = \frac{1}{n} \sum_{i=1}^n a_i \sigma(\omega_i^* \cdot \tilde{x} + b_i^*).$$

where

$$a_i = \frac{\gamma(f) \kappa(\theta_i^*)}{(1 + |\omega_i^*|)^m} \leq \frac{\gamma(f)}{2\pi |\hat{\sigma}(a)|}$$

Noting that $\frac{1}{1+|\omega|} \leq 1$, we have

$$\|D^\alpha g\|_\infty = \max_{0 \leq |\alpha| \leq m} \max_{\theta \in \mathbb{R}^d, \theta \in \mathbb{R}^{d+1}} |\kappa(\theta)| \frac{|D^\alpha \sigma(\omega \cdot \tilde{x} + b)|}{(1 + |\omega|)^m} \leq \frac{\|\sigma\|_{m,\infty}}{2\pi|\hat{\sigma}(a)|}$$

Then we have:

$$\begin{aligned} \gamma(f) &= |a|^d \int_{\mathbb{R}^d} (1 + |\omega|)^m \int_{-(1+|\omega|)M}^{(1+|\omega|)M} db |\hat{f}(a\omega)| d\omega \\ (20.25) \quad &= |a|^d \int_{\mathbb{R}^d} 2M(1 + |\omega|)^{m+1} |\hat{f}(a\omega)| d\omega \\ &\leq 2M \int_{\mathbb{R}^d} (1 + |\omega/a|)^{m+1} |\hat{f}(\omega)| d(\omega) \end{aligned}$$

Denote

$$\|f\|_{m+1} = \int_{\mathbb{R}^d} (1 + |\omega/a|)^{m+1} |\hat{f}(\omega)| d(\omega)$$

Then there exists $(\omega_i^*, b_i^*)_{i=1}^n$ and β_i such that

$$\|f - f_n\|_m \leq \frac{\|\sigma\|_{m,\infty}}{\sqrt{n}} \|f\|_{m+1}$$

where $C = \frac{M\sqrt{C_{m+d}^m}}{\pi|\hat{\sigma}(a)|}$.

20.2.2 Periodic activation function

Now σ is periodic with period T . Then we can write the Fourier series for σ :

$$\sigma(x) = \sum_{i=-\infty}^{\infty} c_n e^{i\frac{2\pi n x}{T}}$$

choose n_1 such that $c_{n_1} \neq 0$, we also have:

$$c_{n_1} = \frac{1}{T} \int_0^T \sigma(x) e^{-i\frac{2\pi n_1 x}{T}} dx = |c_{n_1}| e^{i\beta_2(c_{n_1})}$$

Then for f :

$$\begin{aligned} f(x) &= \int_{\mathbb{R}^d} e^{i\omega \cdot x} e^{i\beta_1(\omega)} |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \frac{1}{c_{n_1}} \frac{1}{T} \int_{x_0}^{x_0+T} \sigma(t) e^{-i\frac{2\pi n_1 t}{T}} dt e^{i(\omega \cdot x) + i\beta_1(\omega)} |\hat{f}(\omega)| d\omega \\ (20.26) \quad &= \int_{\mathbb{R}^d} \frac{1}{|c_{n_1}| e^{i\beta_2(c_{n_1})}} \frac{1}{2\pi n_1} \int_0^{2\pi n_1} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) e^{-i(\omega \cdot x + b)} db e^{i(\omega \cdot x) + i\beta_1(\omega)} |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \frac{1}{2|c_{n_1}| \pi n_1} \int_0^{2\pi n_1} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) e^{i(\beta_1(\omega) - b - \beta_2(c_{n_1}))} db |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} \int_0^{2\pi n_1} \kappa(b, \omega) \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) db |\hat{f}(\omega)| d\omega \end{aligned}$$

here we do a substitution $t = \frac{T}{2\pi n_1}(\omega \cdot x + b)$ and

$$(20.27) \quad \kappa = \frac{\cos(\beta_1(\omega) - b - \beta_2(c_{n_1}))}{2|c_{n_1}|\pi n_1}$$

which is then the same as what we did for compactly supported activation functions.

$$(20.28) \quad \begin{aligned} f(x) &= \int_{\mathbb{R}^d} \int_0^{2\pi n_1} \kappa(b, \omega) \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) db |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d \times [0, 2\pi n_1]} \frac{\kappa(\theta)}{(1 + |\omega|)^m} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) (1 + |\omega|)^m |\hat{f}(\omega)| d\theta \\ &= \int_{\mathbb{R}^d \times [0, 2\pi n_1]} \frac{\kappa(\theta)\gamma(f)}{(1 + |\omega|)^m} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) \frac{(1 + |\omega|)^m |\hat{f}(\omega)|}{\gamma(f)} d\theta \\ &:= \int_{\mathbb{R}^{d+1}} \frac{\kappa(\theta)\gamma(f)}{(1 + |\omega|)^m} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right) d\lambda \end{aligned}$$

where

$$\gamma(f) = \int_{\mathbb{R}^{d+1}} \mathbf{1}_{0 \leq b \leq 2\pi n_1} (1 + |\omega|)^m |\hat{f}(\omega)| d\theta = 2\pi n_1 \int_{\mathbb{R}^d} (1 + |\omega|)^m |\hat{f}(\omega)| d\omega$$

and

$$d\lambda = \frac{(1 + |\omega|)^m |\hat{f}(\omega)| \mathbf{1}_{0 \leq b \leq 2\pi n_1}}{\gamma(f)} d\theta, \quad \int_{\mathbb{R}^{d+1}} d\lambda = 1,$$

We can write

$$(20.29) \quad \tilde{f}(x) \equiv \frac{f(x)}{\gamma(f)} = \mathbb{E}(g(\theta; x))$$

where

$$(20.30) \quad g(\theta; x) = \frac{\kappa(\theta)}{(1 + |\omega|)^m} \sigma\left(\frac{T}{2\pi n_1}(\omega \cdot x + b)\right)$$

And

$$(20.31) \quad D^\alpha \tilde{f}(x) = \mathbb{E}(D^\alpha g).$$

Let $\theta_i = (\omega_i, b_i)_{i=1}^n$ be independently drawn from the same distribution λ and let

$$\bar{g}(\theta_1, \dots, \theta_n) = \frac{1}{n} \sum_{i=1}^n g(\theta_i, x).$$

Then

$$\bar{\mathbb{E}} \left(\sum_{|\alpha| \leq m} (D^\alpha \tilde{f} - D^\alpha \bar{g})^2 \right) = \sum_{|\alpha| \leq m} \bar{\mathbb{E}}(D^\alpha \tilde{f} - D^\alpha \bar{g})^2 = \sum_{|\alpha| \leq m} \bar{\mathbb{E}}(\mathbb{E}(D^\alpha g) - D^\alpha \bar{g})^2 \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

Taking the L^2 norm (with a probability measure) on both side of the above inequality and using Fubini's theorem,

$$(20.32) \quad \mathbb{E}(\|\tilde{f}(\cdot) - g(\theta, \cdot)\|_{H^m(\Omega)}^2) \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

Thus there exists $\theta_i = (\omega_i^*, b_i^*)_{i=1}^n$ and β_i such that

$$\|\tilde{f} - \frac{1}{n} \sum_{i=1}^n g(\theta_i; \cdot)\|_m^2 \leq \frac{1}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

$$\|f - f_n\|_m^2 \leq \frac{\gamma(f)^2}{n} \sum_{|\alpha| \leq m} \|D^\alpha g\|_\infty^2.$$

where

$$f_n(x) = \frac{\gamma(f)}{n} \sum_{i=1}^n g(\theta_i; \cdot) = \frac{1}{n} \sum_{i=1}^n a_i \sigma\left(\frac{T}{2\pi n_1}(\omega_i^* \cdot x + b_i^*)\right).$$

where

$$a_i = \frac{\gamma(f)\kappa(\theta_i^*)}{(1 + |\omega_i^*|)^m} \leq \frac{\gamma(f)}{2|c_{n_1}|\pi n_1}$$

Noting that $\frac{1}{1+|\omega|} \leq 1$, we have

$$\|D^\alpha g\|_\infty = \max_{0 \leq |\alpha| \leq m} \max_{\theta \in \mathbb{R}^d, \theta \in \mathbb{R}^{d+1}} |\kappa(\theta)| \frac{|D^\alpha \sigma(\frac{T}{2\pi n_1}(\omega \cdot x + b))|}{(1 + |\omega|)^m} \leq \frac{k\|\sigma\|_{m,\infty}}{2|c_{n_1}|\pi n_1}$$

where

$$k = \max\left(\frac{T}{2\pi n_1}, \left(\frac{T}{2\pi n_1}\right)^\alpha\right)$$

Denote

$$\|f\|_m = \int_{\mathbb{R}^d} (1 + |\omega|)^m |\hat{f}(\omega)| d(\omega)$$

Then there exists $(\omega_i^*, b_i^*)_{i=1}^n$ and β_i such that

$$\|f - f_n\|_m \leq \frac{\|\sigma\|_{m,\infty}}{\sqrt{n}} \|f\|_m$$

where $C = \frac{k\sqrt{C_{m+d}^m}}{2|c_{n_1}|\pi n_1}$.

Exponential Decay

If $\|e^{\eta|t|} D^k \sigma(t)\|_{L^1(\mathbb{R})}$ is finite for some $\eta > 0$ and all $0 \leq k \leq m$, say

$$\|e^{\eta|t|} D^k \sigma(t)\|_{L^1(\mathbb{R})} \leq \zeta < \infty,$$

which means:

$$\|D^k \sigma(t)\|_{L^1(|t|>M)} \leq e^{-\eta M} \zeta,$$

We have

$$\begin{aligned}
 (20.33) \quad f(x) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}} \kappa(\omega, b) \sigma(\omega \cdot x + b) |a|^d |\hat{f}(a\omega)| db d\omega. \\
 &= \int_{\mathbb{R}^d} \left(\int_{|b| \leq (1+|\omega|)M} + \int_{|b| > (1+|\omega|)M} \right) \frac{\kappa(\omega, b)}{(1+|\omega|)^m} \sigma(\omega \cdot x + b) (1+|\omega|)^m |a|^d |\hat{f}(a\omega)| db d\omega. \\
 &:= f_M + f_r
 \end{aligned}$$

For f_M , from the previous result, we know there exists f_n such that

$$(20.34) \quad \|f_M - f_n\|_m \leq C \frac{\|\sigma\|_{m,\infty}}{\sqrt{n}} \|f\|_{m+1}$$

where $C = \frac{M \sqrt{C_{m+d}^m}}{\pi |\hat{\sigma}(a)|}$ and

$$f_n(x) = \sum_{i=1}^n a_i \sigma(\omega_i^* \cdot \tilde{x} + b_i^*).$$

with

$$a_i = \frac{\gamma(f) \kappa(\theta_i^*)}{n(1+|\omega_i^*|)}$$

As for f_r :

$$\begin{aligned}
 (20.35) \quad |D^\alpha f_r| &\leq \frac{1}{2\pi |\hat{\sigma}(a)|} \|D^{|\alpha|} \sigma(t)\|_{L^1(|t| \geq M)} \int_{\mathbb{R}^d} (1+|\omega/a|)^m |\hat{f}(\omega)| d(\omega) \\
 &= \frac{1}{2\pi |\hat{\sigma}(a)|} \|D^{|\alpha|} \sigma(t)\|_{L^1(|t| \geq M)} \|f\|_m \\
 &\leq \frac{1}{2\pi |\hat{\sigma}(a)|} \|D^{|\alpha|} \sigma(t)\|_{L^1(|t| \geq M)} \|f\|_{m+1} \\
 &\leq \frac{1}{2\pi |\hat{\sigma}(a)|} e^{-\eta M} \zeta \|f\|_{m+1}
 \end{aligned}$$

Thus take the L^2 norm w.r.t a probability measure on Ω and sum over $0 \leq |\alpha| \leq m$, we have:

$$(20.36) \quad \|f_r\|_m^2 \leq \frac{C_{m+d}^m}{(2\pi |\hat{\sigma}(a)|)^2} e^{-2\eta M} \|f\|_{m+1}^2$$

Choose $M = \frac{\ln n}{2\eta}$, then $e^{-\eta M} = \frac{1}{\sqrt{n}}$. Combine (20.34) and (20.36), we get:

$$\begin{aligned}
 (20.37) \quad \|f - f_n\|_m &\leq \|f_r\|_m^2 + \|f_M - f_n\|_m \\
 &\leq \frac{\sqrt{C_{m+d}^m}}{2\pi |\hat{\sigma}(a)|} \frac{1}{\sqrt{n}} \|f\|_{m+1} + \frac{\sqrt{C_{m+d}^m}}{2\eta\pi |\hat{\sigma}(a)|} \frac{\|\sigma\|_{m,\infty} \ln n}{\sqrt{n}} \|f\|_{m+1} \\
 &\leq \frac{\sqrt{C_{m+d}^m}}{2\pi |\hat{\sigma}(a)|} \|f\|_{m+1} \left(\frac{1}{\sqrt{n}} + \frac{\ln n \|\sigma\|_{m,\infty}}{\eta \sqrt{n}} \right)
 \end{aligned}$$

20.3 Barron theory of approximation

20.3.1 Approximation result for cosine as activation function

For clarity, let us give a brief introduction of the approximation result based on the results in Jones [?], Barron [?] and some modification in Xu [?].

Given d and n , consider the following nonlinear space in \mathbb{R}^d :

$$(20.38) \quad V_n = \left\{ v : v(x) = \sum_{j=1}^n \frac{a_j}{n} \cos(\omega_j \cdot x + b_j) + c, \text{ with } a_j, b_j, c \in \mathbb{R}, \omega_j \in \mathbb{R}^d \right\}.$$

Theorem 126. *Given a bounded domain $B \subset \mathbb{R}^d$, a probability measure μ on B and a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ whose Fourier transform \hat{f} satisfying $\|\hat{f}\|_{L^1(\mathbb{R}^d)} < \infty$. Then, for any $n \geq 1$, there exists $f_n \in V_n$ such that*

$$(20.39) \quad \|f - f_n\|_{L^2(B)} \leq \frac{2\|\hat{f}\|_{L^1(\mathbb{R}^d)}}{\sqrt{n}},$$

where $f_n(x) = \sum_{j=1}^n \frac{a_j}{n} \cos(\omega_j \cdot x + b_j) + c$ satisfying

$$(20.40) \quad a = \|\hat{f}\|_{L^1(\mathbb{R}^d)}, \quad |b_j| \leq \pi, \quad |c| \leq \|f\|_{0,\infty,B} + \|\hat{f}\|_{L^1(\mathbb{R}^d)}.$$

Proof. Consider the Fourier transform:

$$(20.41) \quad \hat{f}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\omega \cdot x} f(x) dx \quad \forall \omega \in \mathbb{R}^d.$$

We write $\hat{f}(\omega) = e^{i\theta(\omega)} |\hat{f}(\omega)|$. By Fourier inversion formula,

$$(20.42) \quad f(x) = \int_{\mathbb{R}^d} e^{i\omega \cdot x} \hat{f}(\omega) d\omega = \int_{\mathbb{R}^d} e^{i(\omega \cdot x + \theta(\omega))} |\hat{f}(\omega)| d\omega.$$

Since $f(x)$ is real-valued, it implies that, for $x, x_B \in B$

$$(20.43) \quad \begin{aligned} f(x) - f(x_B) &= \operatorname{Re} \int_{\mathbb{R}^d} (e^{i\omega \cdot x} - e^{i\omega \cdot x_B}) \hat{f}(\omega) d\omega \\ &= \operatorname{Re} \int_{\mathbb{R}^d} (e^{i\omega \cdot x} - e^{i\omega \cdot x_B}) e^{i\theta(\omega)} |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} (\cos(\omega \cdot x + \theta(\omega)) - \cos(\omega \cdot x_B + \theta(\omega))) |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} (\cos(\omega \cdot (x - x_B) + \theta_B(\omega)) - \cos(\theta_B(\omega))) |\hat{f}(\omega)| d\omega \\ &= \int_{\mathbb{R}^d} g(x, \omega) |\hat{f}(\omega)| d\omega \\ &= \|f\|_{B^m} \int_{\mathbb{R}^d} |\omega|_B^{-m} g(x, \omega) \lambda^m(\omega) d\omega \end{aligned}$$

where

$$\theta_B(\omega) = \omega \cdot x_B + \theta(\omega), \quad |\omega|_B := \sup_{x \in B} |\omega \cdot (x - x_B)|$$

and $g : B \times \mathbb{R}^d \rightarrow \mathbb{R}$ is given by

$$(20.44) \quad g(x, \omega) := \cos(\omega \cdot (x - x_B) + \theta_B(\omega)) - \cos(\theta_B(\omega)),$$

and

$$\|f\|_{B^m} := \int_{\mathbb{R}^d} |\omega|_B^m |\hat{f}(\omega)| d\omega, \quad \lambda^m(\omega) = \frac{|\omega|_B^m |\hat{f}(\omega)|}{\|f\|_{B^m}}.$$

here λ^m is a probability distribution density function.

Using $\lambda = \lambda^0$, we define the expectation \mathbb{E} and $\bar{\mathbb{E}}$ for $u : \mathbb{R}^d \mapsto \mathbb{R}$ and $v : (\mathbb{R}^d)^n \equiv \mathbb{R}^d \times \cdots \times \mathbb{R}^d \mapsto \mathbb{R}$

$$\mathbb{E}(u) \equiv \int_{\mathbb{R}^d} u(\omega) \lambda(\omega) d\omega, \quad \bar{\mathbb{E}}(v) \equiv \int_{(\mathbb{R}^d)^n} v(\omega_1, \dots, \omega_n) \lambda(\omega_1) \cdots \lambda(\omega_n) d\omega_1 \cdots d\omega_n.$$

Denote $\tilde{f}(x) = \frac{f(x) - f(x_B)}{\|f\|_B}$, by Lemma ?? and a direct (and standard) calculation

$$\bar{\mathbb{E}}(\tilde{f}(x) - \frac{1}{n} \sum_{j=1}^n g(x, \omega_j))^2 = \bar{\mathbb{E}}(\mathbb{E}[g(x, \cdot)]) - \frac{1}{n} \sum_{j=1}^n g(x, \omega_j)^2 \leq \frac{1}{n} \max_{x \in B, \omega \in \mathbb{R}^d} |g(x, \omega)|^2 \leq \frac{4}{n}.$$

Integrating both sides on B , then by Fubini's Theorem we should have

$$\bar{\mathbb{E}} \int_B (\tilde{f}(x) - \frac{1}{n} \sum_{j=1}^n g(x, \omega_j))^2 d\mu(x) \leq \frac{4}{n}.$$

Thus $\exists \omega_j^* \in \mathbb{R}^d$ such that

$$\int_B (\tilde{f}(x) - \frac{1}{n} \sum_{j=1}^n g(x, \omega_j^*))^2 d\mu(x) \leq \frac{4}{n}.$$

The desired result then follows easily. \square

To prove an H^1 estimate, we use the following density function

$$\lambda_1(\omega) = \lambda^1 = \frac{|\omega|_B |\hat{f}(\omega)|}{\int_{\mathbb{R}^d} |\omega|_B |\hat{f}(\omega)|}.$$

Similarly, we have

$$\tilde{f}(x) \equiv \frac{f(x) - f(x_B)}{\int_{\mathbb{R}^d} |\omega|_B |\hat{f}(\omega)|} = \int_{\mathbb{R}^d} \frac{g(x, \omega)}{|\omega|_B} \lambda_1(\omega) d\omega.$$

It follows that

$$\partial_k \tilde{f}(x) = \int_{\mathbb{R}^d} \frac{\partial_k g(x, \omega)}{|\omega|_B} \lambda_1(\omega) d\omega.$$

We note that

$$\max_{x \in B, \omega \in \mathbb{R}^d} \frac{|g(x, \omega)|}{|\omega|_B} \leq 1.$$

Also, by the definition of $|\omega|_B$, we know there exists $x_\omega \in B$ such that:

$$|\omega|_B = \sup_{x \in B} |\omega \cdot (x - x_B)| \geq |\omega| |x_\omega - x_B| \geq |\omega| \frac{1}{2} \text{dist}(x_B, \partial B).$$

Thus

$$\max_{x \in B, \omega \in \mathbb{R}^d} \frac{|\partial_k g(x, \omega)|}{|\omega|_B} \leq \frac{2}{\text{dist}(x_B, \partial B)}.$$

Setting

$$v_n = \frac{1}{n} \sum_{j=1}^n \frac{g(x, \omega_j)}{|\omega_j|_B}.$$

By a similar argument, we obtain

$$\mathbb{E} \int_B \left((\tilde{f}(x) - v_n(x))^2 + \sum_{k=1}^d (\partial_k \tilde{f}(x) - \partial_k v_n(x))^2 \right) d\mu(x) \leq \frac{1}{n} [C(d, B)]^2.$$

where $C(d, B) = [\frac{4d}{\text{dist}^2(x_B, \partial B)} + 1]^{1/2}$. This implies that $\exists \omega_j^* \in \mathbb{R}^d$ such that

$$\left\| \tilde{f}(\cdot) - \frac{1}{n} \sum_{j=1}^n \frac{g(\cdot, \omega_j^*)}{|\omega_j^*|_B} \right\|_{1, B}^2 \leq \frac{[C(d, B)]^2}{n}.$$

Theorem 127. *Given a bounded domain $B \subset \mathbb{R}^d$, a probability measure μ on B and a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ whose Fourier transform \hat{f} satisfying $\|f\|_{B^1} < \infty$. Then, for any $n \geq 1$, there exists $f_n \in V_n$ such that*

$$(20.45) \quad \|f - f_n\|_{H^1(B)} \leq \frac{C(d, B)}{\sqrt{n}} \|f\|_{B^1}.$$

where

$$(20.46) \quad f_n(x) = \sum_{j=1}^n a_j \cos(\omega_j \cdot x + b_j) + c \text{ for some } a_j, b_j, c \in \mathbb{R}, \omega_j \in \mathbb{R}^d.$$

One remarkable fact is that Theorem 127 holds in any spatial dimension and any bounded domain B of any geometric shape. Theorem 127 is only interesting when $d \geq 2$. For $d = 1$, we can prove much stronger results easily. For example, if u is sufficiently smooth, for any $m \geq 1$, we can find functions u_n in the form of (20.46) such that

$$(20.47) \quad \|u - u_n\|_{1, B} \leq \frac{C_1(m, u)}{n^m}.$$

20.4 An improved analysis

20.4.1 Heaviside Function

Define $g_i : [-1, 1] \mapsto \mathbb{R}$ as follows:

$$(20.48) \quad g_i(t) = \frac{1}{|\omega_i|_B} [\cos(|\omega_i|_B t + \theta_B(\omega_i)) - \cos(\theta_B(\omega_i))],$$

In view of (20.44), we have

$$(20.49) \quad g_i(s_i) = \frac{g(x, \omega_i)}{|\omega_i|_B}, \quad s_i = \omega_i^B \cdot (x - x_B), \quad \omega^B = \frac{\omega}{|\omega|_B}$$

Now, we take an integer

$$(20.50) \quad k \geq \sqrt{n}$$

and consider a partition of $[-1, 1]$ with the following grid points

$$t_j = jh_k, \quad j = -k : k$$

with

$$h_k = \frac{1}{k} \leq \frac{1}{\sqrt{n}}.$$

We first take a piecewise constant interpolation for g_i on $[0, 1]$ to get

$$g_{i,k}(t) = (I_k g_i)(t) = \sum_{j=0}^{k-1} g_i(t_j) M_j(t),$$

where

$$M_j(t) = M_0\left(\frac{t - t_j}{h_k}\right)$$

and

$$(20.51) \quad M_0(x) = \begin{cases} 0 & x \leq 0 \\ 1 & 0 < x \leq 1 \\ 0 & x > 1 \end{cases}$$

We note that

$$M_0(x) = H(x) - H(x - 1)$$

where H is the Heaviside function Thus

$$M_0\left(\frac{t - t_j}{h_k}\right) = H\left(\frac{t - t_j}{h_k}\right) - H\left(\frac{t - t_j}{h_k} - 1\right) = H\left(\frac{t - t_j}{h_k}\right) - H\left(\frac{t - t_{j+1}}{h_k}\right) \equiv H_j(t) - H_{j+1}(t).$$

Thus, since $g_i(t_0) = 0, H_k = 0$, we have

$$(20.52) \quad g_{i,k}(t) = \sum_{j=0}^{k-1} g_i(t_j) M_j(t) = \sum_{j=1}^{k-1} (g_i(t_j) - g_i(t_{j-1})) H_j(t), \quad t \in [0, 1]$$

Now we consider

$$(20.53) \quad h_i(t) = g_i(-t), \quad t \in [0, 1].$$

Similar to (20.52), we have

$$(\Pi_k h_i)(t) = \sum_{j=1}^{k-1} (h_i(t_j) - h_i(t_{j-1}))H_j(t) = \sum_{j=1}^{k-1} (g_i(-t_j) - g_i(-t_{j-1}))H_j(t)$$

Namely

$$(\Pi_k g_i)(-t) = \sum_{j=1}^{k-1} (g_i(-t_j) - g_i(-t_{j-1}))H_j(t), \quad t \in [0, 1]$$

or

$$(20.54) \quad (\Pi_k g_i)(t) = \sum_{j=1}^{k-1} (g_i(-t_j) - g_i(-t_{j-1}))H_j(-t), \quad t \in [-1, 0]$$

By combining (20.52) and (20.65), we get a piecewise constant interpolation of g_i on $[-1, 1]$ as follows:

$$(20.55) \quad \begin{aligned} g_{i,k}(t) &= \sum_{j=1}^{k-1} (g_i(-t_j) - g_i(-t_{j-1}))H_j(-t) + \sum_{j=1}^{k-1} (g_i(t_j) - g_i(t_{j-1}))H_j(t) \\ &= \sum_{j=1}^{k-1} [a_{ij}^- H_j(-t) + a_{ij}^+ H_j(t)] \quad t \in [-1, 1] \end{aligned}$$

where

$$a_{ij}^\pm = g_i(\pm t_j) - g_i(\pm t_{j-1})$$

It is easy to see that

$$(20.56) \quad |g_i(t) - g_{i,k}(t)| \leq h_k, \quad t \in [-1, 1].$$

$$(20.57) \quad \left\| \frac{1}{n} \sum_{i=1}^n \frac{g(\cdot, \omega_i)}{|\omega_i|_B} - f^* \right\|_{L^2(\mu, B)} \leq h_k$$

where

$$(20.58) \quad f^*(x) = \frac{1}{n} \sum_{i=1}^n g_{i,k}(\omega_i^B \cdot (x - x_B)).$$

By the approximation in last section, we have

$$(20.59) \quad \|\tilde{f} - f^*\|_{L^2(\mu, B)} \leq \frac{2}{\sqrt{n}}$$

Let us rewrite

$$f^*(x) = \sum_{i=1}^n \sum_{j=1}^{k-1} [\gamma_{ij}^- f_{ij}^- + \gamma_{ij}^+ f_{ij}^+]$$

where

$$\gamma_{ij}^{\pm} = \frac{|a_{ij}^{\pm}|}{nd_i}, f_{i,j}^{\pm} = d_i \text{sign}(a_{ij}^{\pm}) H_j(\pm \omega_i^B \cdot (x - x_B))$$

and

$$d_i = \sum_{j=1}^{k-1} (|a_{ij}^-| + |a_{ij}^+|) \leq 2$$

By definition

$$(20.60) \quad \sum_{i=1}^n \sum_{j=1}^{k-1} [\gamma_{i,j}^- + \gamma_{i,j}^+] = 1.$$

With re-numeration as

$$p_{\ell} = \gamma_{ij}^{\pm}, f_{\ell} = f_{ij}^{\pm}, 1 \leq \ell \leq N = 2n(k-1)$$

We have

$$f^*(x) = \sum_{\ell=1}^N p_{\ell} f_{\ell}$$

Consider

$$\mathcal{N} = \{1, 2, \dots, N\}$$

and

$$\bar{f} : \mathcal{N} \mapsto \mathbb{R}^1$$

such that

$$\bar{f}(\ell) = f_{\ell}, \ell \in \mathcal{N}$$

With the probability measure

$$\mu(\mathcal{M}) = \sum_{m \in \mathcal{M}} p_m \quad \mathcal{M} \subset \mathcal{N}.$$

By definition.

$$\mathbb{E}(\bar{f}) = f^*(x).$$

By the basic result on expectation in Lemma ??, we have

$$\sum_{\ell_1, \dots, \ell_n=1}^N p_{\ell_1} \cdots p_{\ell_n} \left(f^*(x) - \frac{1}{n} \sum_{i=1}^n f_{\ell_i} \right)^2 = \mathbb{E}_{\mathcal{N}^n} \left(\mathbb{E}(\bar{f}) - \frac{1}{n} \sum_{i=1}^n \bar{f}(\ell_i) \right)^2 \leq \frac{1}{n} \|\bar{f}\|_{\infty}^2 \leq \frac{4}{n}.$$

By taking the $L^2(\mu, B)$ on the above inequality, we get

$$\sum_{\ell_1, \dots, \ell_n=1}^N p_{\ell_1} \cdots p_{\ell_n} \|f^* - \frac{1}{n} \sum_{i=1}^n f_{\ell_i}\|_{L^2(\mu, B)}^2 \leq \frac{4}{n}.$$

Thus, there exist $\ell_1^*, \dots, \ell_n^* \in \mathcal{N}$ such that

$$\|f^* - \frac{1}{n} \sum_{i=1}^n f_{\ell_i^*}\|_{L^2(\mu, B)}^2 \leq \frac{4}{n}.$$

where

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n f_{\ell_i}(x).$$

Then we have

$$(20.61) \quad \|\tilde{f} - f_n\|_{L^2(\mu, B)}^2 \leq \frac{9}{n}.$$

Consequently

$$(20.62) \quad \|f(x) - f(x_B) - \|f\|_B f_n\|_{L^2(\mu, B)}^2 \leq \frac{9\|f\|_B^2}{n}.$$

20.4.2 Piecewise linear function

The proof here is almost the same as the proof for Heaviside function in the last part. Now we take a piecewise linear interpolation for g_i on $[0, 1]$, since $g_i(t_0) = 0$, we get

$$g_{i,k}(t) = (\Pi_k g_i)(t) = \sum_{j=1}^k [g_i(t_j) - g_i(t_{j-1})] \sigma_{j-1}(t), \quad t \in [0, 1]$$

where

$$\sigma_j(t) = M_0\left(\frac{t - t_j}{h_k}\right)$$

and

$$(20.63) \quad M_0(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 < x \leq 1 \\ 1 & x > 1 \end{cases}$$

Consider

$$(20.64) \quad h_i(t) = g_i(-t), \quad t \in [0, 1].$$

Similarly, we have

$$(\Pi_k h_i)(t) = \sum_{j=1}^k (h_i(t_j) - h_i(t_{j-1})) \sigma_{j-1}(t) = \sum_{j=1}^k (g_i(-t_j) - g_i(-t_{j-1})) \sigma_{j-1}(t)$$

Namely

$$(\Pi_k g_i)(-t) = \sum_{j=1}^k (g_i(-t_j) - g_i(-t_{j-1})) \sigma_{j-1}(t), \quad t \in [0, 1]$$

or

$$(20.65) \quad (\Pi_k g_i)(t) = \sum_{j=1}^k (g_i(-t_j) - g_i(-t_{j-1})) \sigma_{j-1}(-t), \quad t \in [-1, 0]$$

Combine together we get a piecewise linear interpolation of g_i on $[-1, 1]$ as follows:

$$\begin{aligned}
 g_{i,k}(t) &= \sum_{j=1}^k (g_i(-t_j) - g_i(-t_{j-1}))\sigma_{j-1}(-t) + \sum_{j=1}^k (g_i(t_j) - g_i(t_{j-1}))\sigma_{j-1}(t) \\
 (20.66) \quad &= \sum_{j=1}^k [a_{ij}^- \sigma_{j-1}(-t) + a_{ij}^+ \sigma_{j-1}(t)] \quad t \in [-1, 1]
 \end{aligned}$$

where

$$a_{ij}^\pm = g_i(\pm t_j) - g_i(\pm t_{j-1})$$

Follow the procedure in last section, and notice that $\sigma(x) = \text{ReLU}(x) - \text{ReLU}(x - 1)$, we obtain the following theorem.

Theorem 128. *For a probability measure μ on B and every function with $\|f\|_B < \infty$, there exists $\omega_1, \dots, \omega_n \in \mathbb{R}^d$ such that*

$$\|f(x) - f_n(x)\|_{L^2(\mu, B)}^2 \leq \frac{C\|f\|_B^2}{n}.$$

where $f_n(x) = \sum_{i=1}^n a_i \text{ReLU}(\omega_i x + b_i) + c$.

20.5 DNN1 versus finite element or wavelets approximation

Three categories of approximation theory.

1. In Barron's paper, there is a section on the lower bound of approximation using linear subspaces. If the basis is fixed, then the rate $n^{-1/d}$ is not improvable. The DNN uses a basis adapt to the function. The adaptive FEM we have studied before is indeed using linear subspaces. For a given and fixed basis, select the best n term to approximate a function. The non-linear approximation theory (by DeVore) is to relax the smoothness of function to achieve the optimal rate $n^{-1/d}$ but won't improve the rate. Now the problem is a truly nonlinear approximation problem, even the basis can be changed according to f . The dimension independent rate $n^{-1/2}$ seems also optimal. What we can improve is the characterization of the smoothness.

20.6 Non-polynomial as activation function

Lemma 128. *Let $\sigma \in C^\infty(\mathbb{R})$ and assume σ is not a polynomial. Then $\Sigma_n(\sigma)$ is dense in $C(\mathbb{R}^n)$.*

Proof. Since $\sigma \in C^\infty(\mathbb{R})$, and $[\sigma((\omega + h e_j) \cdot x + \theta) - \sigma(\omega \cdot x + \theta)]/h \in \Sigma_n(\sigma)$ for every ω, θ and $h \neq 0$, it follows that $\frac{\partial}{\partial \omega_j} \sigma(\omega \cdot x + \theta) \in \overline{\Sigma}_n(\sigma)$ for all $j = 1 : n$. By the same argument $\frac{\partial^k}{\partial \omega_j^k} \sigma(\omega \cdot x + \theta) \in \overline{\Sigma}_n(\sigma)$ for all $k \in \mathbb{N}$, $j = 1 : n$, $\omega \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$.

Now $\frac{\partial^k}{\partial \omega_j^k} \sigma(\omega \cdot x + \theta) = x_j^k \sigma^{(k)}(\omega \cdot x + \theta)$, and since σ is not a polynomial there exists a $\theta_k \in \mathbb{R}$ such that $\sigma^{(k)}(\theta_k) \neq 0$. Take $\omega = 0$ and $\theta = \theta_k$, we then have $x_j^k \in \overline{\Sigma}_n(\sigma)$. Similarly, for all polynomials of the form $x_1^{k_1} \dots x_n^{k_n}$, we can get them by taking the corresponding partial derivatives.

This implies that $\overline{\Sigma}_n(\sigma)$ contains all polynomials. By Weierstrass's Theorem it follows that $\overline{\Sigma}_n(\sigma)$ contains $C(K)$ for each compact $K \subset \mathbb{R}^n$. That is $\Sigma_n(\sigma)$ is dense in $C(\mathbb{R}^n)$. \square

Theorem 129. *Let σ be a non-polynomial Riemann integrable function and $\sigma \in L_{loc}^\infty(\mathbb{R})$. Then $\Sigma_1(\sigma)$ is dense in $C(\mathbb{R})$.*

Proof. Consider the mollifier η

$$\eta(x) = \begin{cases} C \exp\left(\frac{1}{|x|^2 - 1}\right), & |x| < 1, \\ 0, & |x| \geq 1. \end{cases}$$

here C is selected so that $\int_{\mathbb{R}} \eta dx = 1$.

Set $\eta_\epsilon = \frac{1}{\epsilon} \eta\left(\frac{x}{\epsilon}\right)$. Then consider σ_{η_ϵ}

$$(20.67) \quad \sigma_{\eta_\epsilon}(x) := \sigma * \eta_\epsilon(x) = \int_{\mathbb{R}} \sigma(x-y) \eta_\epsilon(y) dy$$

It can be seen that $\sigma_{\eta_\epsilon} \in C^\infty(\mathbb{R})$. Following the proof in the previous proposition, we want to show that $\overline{\Sigma}_1(\sigma)$ contains all polynomials.

The first step is to show that $\overline{\Sigma}_1(\sigma_{\eta_\epsilon}) \subset \overline{\Sigma}_1(\sigma)$, which can be done easily by checking the Riemann sum of $\sigma_{\eta_\epsilon}(x) = \int_{\mathbb{R}} \sigma(x-y) \eta_\epsilon(y) dy$ is in $\overline{\Sigma}_1(\sigma)$.

Then it suffices to show that there exists θ_k and σ_{η_ϵ} such that $\sigma_{\eta_\epsilon}^{(k)}(\theta_k) \neq 0$ for each k . If not, then there must be k_0 such that $\sigma_{\eta_\epsilon}^{(k_0)}(\theta) = 0$ for all $\theta \in \mathbb{R}$ and all $\epsilon > 0$. Thus σ_{η_ϵ} 's are all polynomials with degree at most $k_0 - 1$. In particular, It is known that $\eta_\epsilon \in C_0^\infty(\mathbb{R})$ and $\sigma * \eta_\epsilon$ uniformly converges to σ on compact sets in \mathbb{R} and $\sigma * \eta_\epsilon$'s are all polynomials of degree at most $k_0 - 1$. Polynomials of a fixed degree form a closed linear subspace, therefore σ is also a polynomial of degree at most $k_0 - 1$, which leads to contradiction.

□