Chapter 1

Universal Approximation Theorems

Mathematics is the art of giving the same name to different things.

– Henri Poincaré



HIS chapter will discuss the universal approximation theorems for neural networks. It explains how neural networks of varying widths/depths can fit complex datasets when their parameters are optimized.

1.1 Shallow Neural Network

1.1.1 Qualitative Studies

In this section, we focus on the approximation theorems for the shallow neural networks (also known as single-layer neural networks) with arbitrarily large width.

$$f(x,\theta)=\sum_{i=1}^{n} c_i\sigma(w_i\cdot x+b_i),\quad x\in X\subset\mathbb{R}^d,$$

where the parameters $\theta=\{(c_i,w_i,b_i)\mid i\in[n]\}\in\mathbb{R}^{(d+2)\times n}$. The function σ is called the **nonlinear** activation function. In addition, we assume X is bounded and compact, e.g., the hypercube $[0,1]^d$ or the sphere $\mathbb{S}^{d-1}$.

Definition 1.1.1. *A signed measure on $(X,\mathcal{A})$ ($\mathcal{A}$ being the Borel σ-algera) is a function $\mu:\mathcal{A}\to\mathbb{R}\cup\{\pm\infty\}$ that*

1. *μ can only attain at most one of $\{-\infty,\infty\}$.*
2. *$\mu(\emptyset)=0$.*
3. *Let $\{E_n\}_{n=1}^{\infty}\subset\mathcal{A}$ be a sequence of pairwise disjoint sets, then*

$$\mu\left(\bigcup_n E_n\right)=\sum_n \mu(E_n).$$

Fact 1.1.2 (Hahn Decomposition). *A finite signed measure μ can be written as*

$$\mu=\mu^+-\mu^-$$

where μ^+ and μ^- are two finite non-negative measures.

We then define the Fourier transform on finite signed measures on $X\subset\mathbb{R}^d$, denoted by $\mathcal{M}(X)$.

Theorem 1.1.3. *If $\mu\in\mathcal{M}(X)$, then the Fourier transform $\mathcal{F}:\mathcal{M}(X)\to C_b(\mathbb{R}^d)$*

$$\mathcal{F}(\mu)=\widehat{\mu}(\zeta)=\int_X e^{-ix\cdot\zeta}\mu(dx)$$

Continue on next page ⇓

Continue from last page ⇑

is well-defined, where C_b means the space of bounded continuous functions.

Proof. Use the Hahn decomposition to define the Fourier transform, then apply the dominated convergence theorem to establish continuity. □

Remark 1.1.4. *If $\widehat{\mu}_1 = \widehat{\mu}_2$, by Fubini's theorem and Fourier transform is invertible on $\mathcal{S}$, for any $\psi \in \mathcal{S}$ ($\mathcal{S}$ means Schwartz function class),*

$$\int_X \psi(x)\mu_1(dx) = \int_X \psi(x)\mu_2(dx)$$

Then, μ is orthogonal to $\mathcal{S}$, and $\mathcal{S}$ is dense in $C_0(X)$ (continuous in X, vanish at infinity), by Riesz-Markov theorem, $\mu \in \mathcal{M}(X)$ is unique.

Definition 1.1.5. *Let μ be a **finite signed measure** on the data $X \subset \mathcal{X}$, and $\sigma : \mathbb{R} \to \mathbb{R}$ be the activation function. Then σ is called **discriminatory** with respect to μ if*

$$\int_X \sigma(w \cdot x + b)\mu(dx) = 0,$$

for every $(w, b) \in \mathbb{R}^{d+1}$ implies $\mu \equiv 0$.

Definition 1.1.6. *A continuous function $\sigma : \mathbb{R} \to [0, 1]$ is called **sigmoidal** if*

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

Example 1.1.7. *For instance, the function*

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{1}{2}(1 + \tanh \frac{x}{2})$$

is sigmoidal.

Theorem 1.1.8. *Sigmoidal functions are discriminatory.*

Proof. For sigmoidal functions, the boundedness of σ implies (dominated convergence theorem) that (fix arbitrary w, b, c)

$$\lim_{N \to \infty} \int_X \sigma(N(w \cdot x + b) + c)\mu(dx) = \mu(\{w \cdot x + b > 0\}) + \sigma(c)\mu(\{w \cdot x + b = 0\}).$$

Therefore (by varying c), for any $(w, b) \in \mathbb{R}^{d+1}$,

$$\mu(\{w \cdot x + b > 0\}) = \mu(\{w \cdot x + b = 0\}) = 0.$$

For any continuous function $h : \mathbb{R} \to \mathbb{R}$, one can construct a sequence of simple functions that converges to h pointwise,

$$\sum_i a_i \operatorname{sgn}(w \cdot x + b_i) \to h(w \cdot x).$$

Therefore,

$$\int_X h(w \cdot x)\mu(dx) = 0.$$

It implies $\widehat{\mu} = 0$, thus $\mu = 0$ by uniqueness. □

The following theorem is guaranteed by the Riesz-Markov theorem.

Theorem 1.1.9. *If σ is discriminatory, then the shallow neural networks*

$$f(x, \theta) = \sum_{i=1}^{N} c_i \sigma(w_i \cdot x + b_i)$$

is dense in $C(X)$.

In other words, for any continuous function $h \in C(X)$ and $\varepsilon > 0$, there exists a shallow sigmoidal neural network f that $\|f - h\|_\infty < \varepsilon$.

Remark 1.1.10. *On the other hand, we observe that polynomials of finite degree cannot be discriminatory because polynomials of affine transformation in x are still polynomials.*

Whether a non-polynomial continuous function $\sigma \in C(\mathbb{R})$ is discriminatory? We first consider $\sigma \in C^\infty(\mathbb{R})$. If the signed measure μ is orthogonal to the linear vector space spanned by $\sigma(w \cdot x + b)$, $(w, b) \in \mathbb{R}^{d+1}$. Then μ is orthogonal to

$$\frac{1}{\varepsilon}\left(\sigma((w + \varepsilon e_i) \cdot x + b) - \sigma(w \cdot x + b)\right)$$

for an arbitrary $\varepsilon > 0$ and $\{e_i\}_{i=1}^d$ denotes the canonical basis in $\mathbb{R}^d$. Thus, μ is orthogonal to linear functions

$$x_i \sigma'(w \cdot x + b)|_{w=0} = x_i \sigma'(b), \quad i \in [d].$$

Similarly, one can take higher derivatives in w along different directions and show

μ is orthogonal to

$$\prod_{k=1}^{n} x_{i_k} \cdot \sigma^{(n)}(w \cdot x + b)|_{w=0} = \left(\prod_{k=1}^{n} x_{i_k} \right) \cdot \sigma^{(n)}(b), \quad i_k \in [d].$$

Therefore, μ is orthogonal to all polynomials, which are dense in $C(X)$. We only need the following fact to guarantee all polynomials are included.

Lemma 1.1.11. *If $\sigma \in C^\infty(\mathbb{R})$ is not a polynomial, then for any $n \in \mathbb{N}$, there exists $b_n \in \mathbb{R}$ that $\sigma^{(n)}(b_n) \neq 0$.*

The proof is trivial. Next, we consider $\sigma \in C(\mathbb{R})$ only. We use an arbitrary mollifier $\phi \in C_c^\infty$ to define σ_ϕ as follows.

$$\sigma_\phi(x) := \int_{\mathbb{R}} \sigma(x - y)\phi(y)dy = \int_{\mathbb{R}} \phi(x - y)\sigma(y)dy.$$

Then

$$\sigma_\phi(w \cdot x + b) = \int_{\mathbb{R}} \sigma(w \cdot x + b - y)\phi(y)dy.$$

The linearity implies μ is orthogonal to $\sigma_\phi(w \cdot x + b) \in C^\infty(\mathbb{R})$. Following the previous strategy, we only need to show that all polynomials can be recovered by differentiating σ_ϕ with respect to w.

Otherwise, σ_ϕ must be a polynomial for every ϕ. We consider the highest degree

$$D := \sup_{\phi \in C_c^\infty(\mathbb{R})} \deg(\sigma_\phi).$$

1. If $D = \infty$, then all polynomials can be recovered by taking a maximizing sequence of ϕ.
2. If $D < \infty$, then we take a sequence of ϕ such that $\sigma_\phi \to \sigma$ uniformly on any compact set, which means σ must be a polynomial itself, contradiction.

This also proves a result in Hornik et al. (1989). The original proof relies on the Stone-Weierstrass theorem because neural networks form an algebra.

Theorem 1.1.12. *The shallow neural networks of the form*

$$f(x, \theta) = \sum_{i=1}^{N} c_i \prod_{j=1}^{J_i} \sigma(w_{ij} \cdot x + b_{ij})$$

Continue on next page ⇓

Continue from last page ⇑

are dense in $C(X)$.

The previous theorems are under the $C(X)$ topology. Next, we briefly consider the $L^2(X, \nu)$ topology.

Definition 1.1.13. *Suppose the function σ satisfies*

$$\sigma(w \cdot x + b) \in L^2(X, \nu)$$

*for all $(w, b) \in \mathbb{R}^d$. Then σ is **discriminatory** in $L^2(X, \nu)$ if for $f \in L^2(X, \nu)$*

$$\int_X \sigma(w \cdot x + b) f(x) \nu(dx) = 0$$

for all $(w, b) \in \mathbb{R}^{d+1}$ implies $f = 0$.

The basic Riesz representation theorem applies here because $L^2(X, \nu)$ is a Hilbert space.

Theorem 1.1.14. *If σ is discriminatory in $L^2(X, \nu)$, then the shallow neural networks*

$$f(x, \theta) = \sum_{i=1}^{N} c_i \sigma(w_i \cdot x + b_i)$$

are dense in $L^2(X, \nu)$.

Proof. The span by $f(x, \theta)$ forms a linear subspace $V \subset L^2(X, \nu)$. If V is not dense in $L^2(X, \nu)$, there is a linear functional $\psi : L^2(X, \nu) \to \mathbb{R}$ such that $\psi(V) = 0$ and $\psi \neq 0$ otherwise. The Riesz representation theorem implies

$$\psi(h) = \int_X g_\psi(x) h(x) \nu(dx).$$

for some nonzero $g_\psi \in (L^2(X, \nu))^* = L^2(X, \nu)$ (reflective). Then by definition,

$$\int_X g_\psi(x) \sigma(w \cdot x + b) \nu(dx) = 0$$

for all $(w, b) \in \mathbb{R}^{d+1}$. Therefore, $g_\psi = 0$, contradiction. □

Example 1.1.15. *The indicator function of half-planes*

$$\sigma(w \cdot x + b) = \operatorname{sgn}(w \cdot x + b)$$

are discriminatory in $L^2(X, \nu)$ (since they are dense in stronger topology).

Example 1.1.16. *The functions $\sin(w \cdot x + b)$ are discriminatory in $L^2(X, \nu)$.*

1.1.2 Quantitative Studies

In this section, we quantify the approximation error of the shallow neural network in the simplest setting. Regarding the approximation error, let h be a function in a specific set $\mathcal{H}$, and the approximation error means we would like to find out the error as follows.

$$\inf_{\theta} \left\| h - \sum_{i=1}^{n} c_i \sigma(w_i \cdot x + b_i) \right\|. \tag{1.1}$$

Common choices for the norm $\|\cdot\|$ can be $C(X)$, $L^2(X)$, etc.

Remark 1.1.17. *Note that for a different $h \in \mathcal{H}$, it allows different parameters θ. Consider another type of approximation error as follows.*

$$\inf_{(w_i, b_i)} \sup_{h \in \mathcal{H}} \inf_{c_i} \left\| h - \sum_{i=1}^{n} c_i \sigma(w_i \cdot x + b_i) \right\|. \tag{1.2}$$

Then, the approximation error (1.1) is no less than (1.2). The error (1.2) is similar to the Kolmogorov n-width.

In general, the rougher the objective function, the more difficult it is to approximate with activation functions. We first introduce a classical result by Andrew Barron, as presented in Barron (1993, 1994). For simplicity, we assume $X = B_1(\mathbf{0})$ as the unit ball centered at the origin.

Definition 1.1.18 (Barron Space). *The Barron space $\mathcal{B}(X)$ is defined by*

$$\mathcal{B}(X) = \left\{ f \in L^1(X) \mid \int_{\mathbb{R}^d} \|\zeta\|_2 |\widehat{f}(\zeta)| d\zeta < \infty \right\},$$

where $\widehat{f}$ denotes the Fourier transform of f. We set

$$\|f\|_{\mathcal{B}} = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \|\zeta\|_2 |\widehat{f}(\zeta)| d\zeta$$

Continue on next page ⇓

Continue from last page ⇑

as the norm for $f \in \mathcal{B}(X)$.

Question 1.1.19. *For any $k > \frac{1}{2}(d+1)$,*

$$W^{k,2}(X) \subset \mathcal{B}(X) \subset W^{1,\infty}(X).$$

Definition 1.1.20. *Let $K \subset L^2(X)$ be a bounded set, and let $\mathrm{co}(K)$ denote the* ***convex hull*** *of K.*

$$\mathrm{co}(K) = \left\{ \sum_i a_i k_i \;\middle|\; k_i \in K,\ a_i \geq 0,\ \sum_i a_i \leq 1 \right\}.$$

Definition 1.1.21. *We consider a covering of K by n balls in the L^2 metric. The optimal diameter is defined as follows.*

$$R_n(K) := \inf\{r > 0 \mid K \text{ can be covered by } n \text{ balls of diameter } r\}.$$

Theorem 1.1.22 (Makovoz (1996)). *Let $f \in \mathrm{co}(K)$. There exists a function $\mathcal{G}^*$ in the form of*

$$\mathcal{G}^* = \sum_{j=1}^{2n} a_j k_j, \quad a_j \geq 0, \quad \sum_j a_j \leq 1,$$

such that

$$\|f - \mathcal{G}^*\|_{L^2(X)} \leq \frac{R_n(K) + \varepsilon}{\sqrt{n}},$$

for an arbitrarily small $\varepsilon > 0$.

Proof. If $f \in \mathrm{co}(K)$, then there exists $k_i \in K$, $i = 1, \cdots, m$ such that

$$f = \sum_{i=1}^{m} a_i k_i, \quad a_i \in (0, 1], \quad \sum_{i=1}^{m} a_i \leq 1,$$

The set $\{k_i\}_{i=1}^m$ can be divided into n groups such that each group is contained

in a ball of radius $R_n(K)+\varepsilon$ for any $\varepsilon>0$. Without loss of generality, we can assume $m>n$ and each group is non-empty.

Let those balls be $B_1,\cdots,B_n$, there is a surjective mapping $\pi:[m]\to[n]$ such that $k_i\in B_{\pi(i)}$. Then we construct a random variable $\mathcal{G}$ as follows.

$$\mathcal{G}=\sum_{j=1}^{n}\mathcal{G}_j,$$

where each $\mathcal{G}_j$ is an independent random variable in the following summation form,

$$\mathcal{G}_j=\frac{A_j}{n_j}\sum_{l=1}^{n_j}g_{j,l},\quad A_j=\sum_{i,\pi(i)=j}a_i,\quad n_j=\lfloor nA_j\rfloor+1,$$

where $g_{j,l}$ of the same j are i.i.d. random variables provided that $\pi(l)=j$.

$$\mathbb{P}[g_{j,l}=k_i]=\frac{a_i}{A_j}.$$

Therefore, $\mathbb{E}[\mathcal{G}_j]=f_j:=\sum_{i,\pi(i)=j}a_ik_i$. Now, we compute

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{G}-f\|^2\right]&=\sum_{j=1}^{n}\mathbb{E}\left[\|\mathcal{G}_j-f_j\|^2\right]\\&=\sum_{j=1}^{n}\left(\frac{A_j}{n_j}\right)^2n_j\mathbb{E}\left[\left\|g_{j,l}-\frac{f_j}{A_j}\right\|^2\right]\\&\le\frac{(R_n(K)+\varepsilon)^2}{n}\sum_{j=1}^{n}A_j\le\frac{(R_n(K)+\varepsilon)^2}{n}.\end{aligned}$$

It means there exists a realization $\mathcal{G}^*$ consists of at most $\sum_{j=1}^{n}n_j\le 2n$ terms from K that

$$\|\mathcal{G}^*-f\|_{L^2(X)}\le\frac{(R_n(K)+\varepsilon)}{\sqrt{n}},$$

and ε can be arbitrarily small. □

Remark 1.1.23. *The result holds if $f\in\overline{\mathrm{co}(K)}$, since one can find a sequence $f_j\in\mathrm{co}(K)$ arbitrarily close to f.*

Remark 1.1.24. *In Barron (1993), the set K is treated as a single group, then the form of $2n$ terms can be replaced by n terms, but the $R_n(K)$ is replaced by $R_1(K)$. Therefore, the above result is better.*

Remark 1.1.25. *The key idea is to treat f as an expectation with respect to a probability measure on K. Then, the Monte Carlo error $\mathcal{O}(n^{-1/2})$ is guaranteed for a single*

realization.

Next, we prove the approximation error bound for Barron space.

Theorem 1.1.26. *Let $\mathcal{B}_{1/2} = \{f \in \mathcal{B}(X) \mid \|f\|_{\mathcal{B}} \leq \frac{1}{2}\}$. Suppose σ is a sigmoidal function, then $\forall f \in \mathcal{B}_{1/2}$, there exists a shallow neural network (with a constant shift)*

$$g = \sum_{i=1}^{2n} a_i \sigma(w_i \cdot x + b_i) + f(0)$$

that

$$\|f - g\|_{L^2(X)} \leq C n^{-\frac{d+1}{2d}} + \varepsilon,$$

where C is an absolute constant and $\varepsilon > 0$ is arbitrary.

Proof. Let K be the bounded set in $L^2(X)$:

$$K = \left\{ \pm \sigma(w \cdot x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

When $\lambda \to \infty$, the limit $\sigma(\lambda(w \cdot x + b))$ becomes the Heaviside function σ_h with the jump at $w \cdot x + b = 0$. Define

$$K_h = \left\{ \pm \sigma_h(w \cdot x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

Then, $\overline{\text{co}(K_h)} \subset \overline{\text{co}(K)}$. Due to the Heaviside function's affinity and boundedness of X, we can restrict w and b to a smaller set,

$$K_h = \left\{ \pm \sigma_h(w \cdot x + b) \mid w \in \mathbb{R}^d, \|w\| = 1, |b| \leq 1 \right\}.$$

For a fixed $w \in \mathbb{R}^d$, our first observation is that

$$\left\{ g(w \cdot x) \;\middle|\; g : [-\|w\|_2, \|w\|_2] \to \mathbb{R},\ g(0) = 0, \|g\|_{TV} \leq 1 \right\} \subset \overline{\text{co}(K_h)},$$

where $\|g\|_{TV}$ is the total variation. Therefore, $\forall w, b \in \mathbb{R}^d \times \mathbb{R}$ that $w \neq \mathbf{0}$,

$$\frac{1}{2\|w\|_2}(\cos(w \cdot x + b) - \cos(b)) \in \overline{\text{co}(K_h)}, \quad x \in X. \tag{1.3}$$

Because

$$\begin{aligned} f(x)-f(0) &= \Re\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}(e^{iw\cdot x}-1)\widehat{f}(w)dw \\ &= \Re\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}(e^{iw\cdot x}-1)e^{i\theta(w)}|\widehat{f}(w)|dw \\ &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\left(\cos(w\cdot x+\theta(w))-\cos(\theta(w))\right)|\widehat{f}(w)|dw \\ &= \int_{\mathbb{R}^d}\left[\frac{1}{2\|w\|_2}\left(\cos(w\cdot x+\theta(w))-\cos(\theta(w))\right)\right]\frac{2\|w\|_2}{(2\pi)^{d/2}}|\widehat{f}(w)|dw, \end{aligned}$$

where $\theta(w)=\arg\widehat{f}(w)$. Meanwhile, we notice that

$$\int_{\mathbb{R}^d}\frac{2\|w\|_2}{(2\pi)^{d/2}}|\widehat{f}(w)|dw = 2\|f\|_{\mathcal{B}}\leq 1,$$

therefore, $f(x)-f(0)\in\overline{\text{co}(K_h)}$.

Finally, we estimate $R_n(K_h)$. When $\|w-w'\|\leq\varepsilon^2$ and $|b-b'|\leq\varepsilon^2$,

$$\int_X|\sigma_h(w\cdot x+b)-\sigma_h(w'\cdot x+b')|^2dx\leq C\varepsilon^2$$

for a certain absolute constant C. Therefore, $R_n(K)\leq C'n^{-\frac{1}{2d}}$ for another absolute constant C'. Then, applying the previous Theorem of Makovoz yields the desired approximation error bound. □

Remark 1.1.27. *All arguments can be adapted to a general $L^2(X,\mu)$, which is equipped with a general measure. In fact, we only need a Hilbert space. Thus, the approximation error can be calculated by the empirical loss.*

Remark 1.1.28. *The constant shift can be constructed by an additional term $\sigma(w\cdot x+b)$ with $b\to\infty$. Thus, the conclusion holds if the shallow neural network g consists of $2n+1$ terms.*

Remark 1.1.29. *The error bound of the above theorem is slightly better than that of the original bound in Barron (1993), where the error is $\mathcal{O}(n^{-1/2})$. We should also notice that when $d=1$, the bound $\mathcal{O}(n^{-1})$ is optimal.*

Example 1.1.30. *Let $f(x)=x$ on $[-1,1]$. Let g be a sum of n Heaviside functions with jumps at $x_0,\cdots,x_{n+1}$ (including the endpoints of the domain), then*

$$\|f-g\|^2=\sum_{i=0}^{n}\int_{x_i}^{x_{i+1}}|x-c_i|^2dx\leq\frac{1}{12}\sum_{i=0}^{n}|x_{i+1}-x_i|^3\leq\frac{2}{3}\frac{1}{(n+1)^2}.$$

This means the optimal approximation error for f with a sum of n Heaviside functions is of the order of $\mathcal{O}(n^{-1})$.

Question 1.1.31. *Is the bound $\mathcal{O}(n^{-\frac{d+1}{2d}})$ optimal for high dimensions? It is answered positively in Siegel and Xu (2024).*

More importantly, the error bound in Barron (1993) shows no dependence on the dimension of the domain X. It initiated an unfortunate discussion about **defeating the curse of dimensionality**.

Remark 1.1.32. *The curse of dimensionality refers to the general phenomenon in which high dimensionality makes the problem substantially more difficult. For approximation theory, it usually means the approximation error roughly scales like*

$$(\text{number of terms})^{-\frac{\text{regularity}}{\text{space dimension}}}$$

This upper bound usually can be constructed.

Example 1.1.33. *We consider function $f(x) = f(|x|) = (1+|x|^2)^{-d/2}$, its Fourier transform is the Bessel kernel*

$$\widehat{f}(\zeta) = \underbrace{\frac{2\pi^{d/2}}{\Gamma(d/2)}}_{=|\mathbb{S}^{d-1}|} K_0(|\zeta|).$$

Then the Barron norm of f grows exponentially fast in dimension.

$$\begin{aligned}\|f\|_{\mathcal{B}} &= \frac{1}{(2\pi)^{d/2}}\left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^2 \int_0^\infty r^d K_0(r)dr \\ &= \frac{1}{(2\pi)^{d/2}}\left(\frac{2\pi^{d/2}}{\Gamma(d/2)}\right)^2 2^{d-1}\left(\Gamma\left(\frac{d+1}{2}\right)\right)^2 \\ &= (2\pi)^{d/2}\left(\frac{\Gamma((d+1)/2)}{\Gamma(d/2)}\right)^2.\end{aligned}$$

It means in high dimensions, the Barron space excludes many smooth functions.

Here, we have used the identity (DLMF, 2025, (10.43.19)):

$$\int_0^\infty x^{s-1}K_t(x)ds = 2^{s-2}\Gamma\left(\frac{s+t}{2}\right)\Gamma\left(\frac{s-t}{2}\right), \quad t > s.$$

Question 1.1.34. *Explain why the curse of dimensionality is NOT defeated in Barron's result.*

We note that the probabilistic error bound depends on the prior knowledge of the sampling distribution, while the distribution is inherently high-dimensional. The result establishes only the existence of a shallow neural network, without providing a computationally tractable algorithm. Jones, see Jones (1992), proposes a greedy algorithm with n steps to construct the shallow neural network iteratively as follows.

$$(\alpha', g') = \underset{\alpha\in[0,1],\, g\in\overline{K(\sigma)}}{\arg\min} \|f - (\alpha h_{s-1} + (1-\alpha)g)\|,$$

and generate $h_s = \alpha' h_{s-1} + (1-\alpha')g'$. The algorithm can achieve the desired $\mathcal{O}(n^{-1/2})$ bound, but its search space is large (depends on the dimension).

Question 1.1.35. *Prove this algorithm can achieve the error bound $\mathcal{O}(n^{-1/2})$ after n steps for $f \in \mathcal{B}_{1/2}$.*

1.1.3 Ridge Bounds

In Barron (1993), the approximation theory is specifically for Barron space and the Heaviside activation function σ_h. What if we allow arbitrary activation functions in the Sobolev space? The following set

$$\mathcal{R}_n = \left\{ \sum_{i=1}^{n} k_i(w_i \cdot x) \mid k_i \in C(\mathbb{R}),\ w_i \in \mathbb{R}^d,\ i \in [n] \right\}$$

consists of all n-term **ridge** functions. The following two theorems of Maiorov are optimal. The exponent exhibits the curse of dimensionality.

Theorem 1.1.36 (Maiorov (1999)). *Let $m \geq 1$, $d \geq 2$, and X denotes the unit ball in $\mathbb{R}^d$. There exists a function $f \in W^{m,2}(X)$ that $\|f\|_{W^{m,2}} \leq 1$ and*

$$\inf_{k\in\mathcal{R}_n} \|f - g\|_{L^2(X)} \geq C n^{-\frac{m}{d-1}},$$

where C is independent of f and n.

The proof for the lower bound is technical. Interested readers can read Maiorov (1999) for details.

Theorem 1.1.37 (Maiorov (1999)). *Let $m \geq 1$, $d \geq 2$, and X denotes the unit ball in $\mathbb{R}^d$. For any $f \in W^{m,2}(X)$ that $\|f\|_{W^{m,2}} \leq 1$,*

$$\inf_{g \in \mathcal{R}_n} \|f - g\|_{L^2(X)} \leq C n^{-\frac{m}{d-1}},$$

where C is independent of f and n.

Preparations for Upper Bound

Before the proof of the upper bound, we state a classical theorem for approximation theory.

Theorem 1.1.38 (Canuto and Quarteroni (1982)). *Let $f \in W^{m,2}(X)$ that $\|f\|_{W^{m,2}} \leq 1$. Denote $\mathcal{P}_k$ the linear space of polynomials of degree $\leq k$ in $\mathbb{R}^d$. Then*

$$\inf_{g \in \mathcal{P}_k} \|f - g\|_{L^2(X)} \leq C k^{-m},$$

where C is independent of k and f.

Proof. The extension theorem of Stein allows us to consider X as a cube $[-1, 1]^d$ instead of a ball. Then on X, the orthogonal polynomials are

$$Q_{l_1, \cdots, l_d}(x) = \prod_{i=1}^{d} P_{l_i}(x_i),$$

where P_{l_i} is the normalized Legendre polynomial of order $l_i \in \mathbb{N}$. Each P_{l_i} satisfies

$$\mathcal{A}_i P_{l_i}(x_i) = -l_i(l_i + 1) P_{l_i}, \quad \mathcal{A}_i := \partial_i (1 - x_i^2) \partial_i.$$

The operator $\mathcal{A}_i$ is a self-adjoint operator. Therefore, let $p = m/2$ (assume m is even),

$$(-l_i(l_i + 1))^p \langle f, Q_{l_1, \cdots, l_d} \rangle = \langle f, \mathcal{A}_i^p Q_{l_1, \cdots, l_d} \rangle = \langle \mathcal{A}_i^p f, Q_{l_1, \cdots, l_d} \rangle.$$

The approximation error satisfies

$$\begin{aligned}\inf_{g\in\mathcal{P}_k}\|f-g\|_{L^2(X)}^2 &= \sum_{l_1+\cdots+l_d>k}|\langle f,Q_{l_1,\cdots,l_d}\rangle|^2\\ &\le \sum_{i=1}^d\sum_{l_i>k/d}|\langle f,Q_{l_1,\cdots,l_d}\rangle|^2\\ &= \sum_{i=1}^d\sum_{l_i>k/d}\left(\frac{1}{l_i(l_i+1)}\right)^{2p}|\langle \mathcal{A}_i^p f,Q_{l_1,\cdots,l_d}\rangle|^2\\ &\le \left(\frac{d}{k}\right)^{4p}\sum_{i=1}^d\|\mathcal{A}_i^p f\|_{L^2(X)}^2.\end{aligned}$$

It remains to show $\|\mathcal{A}_i f\|_{L^2(X)}\le C\|f\|_{W^{2,2}(X)}$, which is straightforward. □

Remark 1.1.39. *If m is an odd integer, we can use $(-\mathcal{A}_i)^{1/2}$, which is uniquely determined by the eigendecomposition. And*

$$\langle(-\mathcal{A}_i)^{1/2}f,(-\mathcal{A}_i)^{1/2}f\rangle=\langle-\mathcal{A}_i f,f\rangle\le\|f\|_{W^{1,2}(X)}^2.$$

Remark 1.1.40. *For orthogonal polynomials on the unit ball, they are also eigenfunctions of a Sturm-Liouville type differential operator. Similar results still hold.*

Question 1.1.41. *Let $f\in W^{m,2}(\mathbb{R}^d,\mu)$ with $\mu(dx)=e^{-x^2}dx$ that*

$$\|f\|_{W^{m,2}(\mathbb{R}^d,\mu)}\le 1.$$

Prove

$$\inf_{g\in\mathcal{P}_k}\|f-g\|_{L^2(\mathbb{R}^d,\mu)}\le Ck^{-m/2}.$$

Remark 1.1.42. *The discounted performance for weighted Sobolev spaces on $\mathbb{R}^d$ is not surprising since the constant C is proportional to $(\mathrm{diam}(X))^m$. The nth Hermite polynomials (and Laguerre polynomials as well) are eigenfunctions of eigenvalues $\mathcal{O}(n)$. We should expect a similar error bound (see Guo et al. (2005)) for $W^{m,2}(\mathbb{R}_+^d,\nu)$ that*

$$\|f\|_{W^{m,2}(\mathbb{R}_+^d,\nu)}^2=\sum_{|\alpha|\le m}\int_{\mathbb{R}_+^d}|f^{(\alpha)}(x)|^2x^\alpha e^{-|x|_1}dx.$$

Proof for Upper Bound

The proof uses the duality of differential operators ∂^α to polynomial x^α, where α is a multi-index.

Proof of Theorem 1.1.37. We construct $\mathcal{P}_k$ as the linear space of homogeneous polynomials of degree k in $\mathbb{R}^d$. Then let $n = \dim(\mathcal{P}_k) = \binom{d-1+k}{k} \leq Ck^{d-1}$, where C depends on d, there exists (why?) $w_i \in \mathbb{R}^d$, $i \in [n]$ that

$$\bigcup_{l=0}^{k} \mathcal{P}_l = \left\{ \sum_{i=1}^{n} k_i(w_i \cdot x) \;\middle|\; k_i \in \bigcup_{l=0}^{k} \pi_l \right\} \subset \mathcal{R}_n,$$

where π_l denotes the linear space of homogeneous polynomials of degree k in $\mathbb{R}^1$. Thus,

$$\inf_{g\in\mathcal{R}_n} \|f-g\|_{L^2(X)} \leq \inf_{g\in\bigcup_{l=0}^{k}\mathcal{P}_l} \|f-g\|_{L^2(X)} \leq C'k^{-m} \leq C''n^{-\frac{m}{d-1}}.$$

□

Question 1.1.43. *Complete the above proof.*

Question 1.1.44. *Since the Barron space $\mathcal{B} \supset W^{m,2}$ for $m > \frac{d+1}{2}$, compare the error bounds from the previous section and Maiorov's result.*

The result of Maiorov shows that the approximation error of Ridge functions is bounded by $\mathcal{O}(n^{-m/(d-1)})$ for the Sobolev space $W^{m,2}$. Thus, shallow neural networks of width n with any activation function must inherit the lower bound. But how sharp is the upper bound?

Theorem 1.1.45 (Maiorov and Pinkus (1999)). *There exists $\sigma \in C^\infty$, sigmoidal, strictly increasing, and for any $g \in \mathcal{R}_n$ and $\varepsilon > 0$, there are parameters (c_i, w_i, b_i) that*

$$\left| g(x) - \sum_{i=1}^{n+d+1} c_i \sigma(w_i \cdot x + b_i) \right| < \varepsilon, \quad x \in X.$$

Rough Idea. The trick is to construct a function by joining a countable dense set of continuous functions with disjoint support. We give the main steps.

1. $C[-1,1]$ is separable, so there is a countable dense subset

$$\{u_k\} \subset C^\infty[-1,1],$$

For instance, the polynomial ring $\mathbb{Q}[x]$ by Weierstrass theorem.

2. On the real axis, put the function $\sigma = a_k u_k + b_k\phi + c_k\psi$ on the disjoint intervals I_k, where ϕ and ψ are linear functions supported on two intervals.

3. Choose the parameters a_k, b_k, c_k such that the whole function on $\mathbb{R}$ is increasing. Then fill the gaps between the intervals smoothly.

Then for any function $g \in \mathcal{R}_n$, it consists of n ridge functions g_i, $i = 1, \cdots, n$. Each g_i is made with σ on an interval and two linear functions in $\mathbb{R}^d$. However, there are at most $(d+1)$ linear functions. Therefore, at most we need $\min(3n, n+d+1)$ terms. □

Question 1.1.46. *Complete this proof.*

Since the dimension d is fixed, $n \sim n+d+1$. This means that the activation function σ in the above theorem is sharp for approximation in Sobolev spaces.

Corollary 1.1.47. *Let σ be the activation function in the previous theorem, and define*

$$\mathcal{S} := \left\{ \sum_{i=1}^{n} c_i \sigma(w_i \cdot x + b_i) \mid c_i \in \mathbb{R},\ w_i \in \mathbb{R}^d,\ b_i \in \mathbb{R} \right\}.$$

Then, for any $f \in W^{m,2}(X)$ with bound $\|f\|_{W^{m,2}(X)} \leq 1$,

$$\inf_{g\in\mathcal{S}} \|f-g\|_{L^2(X)} \leq Cn^{-m/(d-1)}.$$

However, such an activation function is too complex (even pathological) to be practical, despite its desirable properties, including smoothness (even analyticity when $n+d+1$ is replaced by $3n$), monotonicity, and sigmoidality. Computationally, simple activation functions are preferred. In return, the approximation error bound could be slightly suboptimal.

Stable Approximation

In this section, we show that the error bound $\mathcal{O}(n^{-m/d})$ is optimal when the approximation is a *continuous* mapping. In this case, we say it is a stable approximation. Note that this is *not* about the representation itself; it is related to how the approximation is found.

Definition 1.1.48 (Nonlinear n-Width). *Let K be a compact set in a normed space $\mathcal{X}$. Let $\mathcal{E}_n : K \to \mathbb{R}^n$ be a continuous map and $\mathcal{D}_n : \mathbb{R}^n \to K$ be a function, unnecessarily continuous. Define*

$$h_n(K;\mathcal{X}) = \inf_{\mathcal{E}_n,\mathcal{D}_n} \sup_{f\in K} \|f - \mathcal{D}_n(\mathcal{E}_n(f))\|_{\mathcal{X}},$$

$h_n(K)$ is the continuous nonlinear n-width.

The following theorem is proved in DeVore et al. (1989).

Theorem 1.1.49 (DeVore et al. (1989)). *Let $m, p, d \geq 1$, and X be the unit ball in $\mathbb{R}^d$. Then*

$$h_n(W^{m,p}(X); L^p(X)) \asymp n^{-m/d}.$$

Proof. The upper bound is not hard, as we can see later. We only show the lower bound. The purpose is to introduce the standard approach.

1. Set (Berstein n-width)

$$b_n(K;\mathcal{X}) := \sup_{\mathcal{X}_{n+1}} \sup\{\lambda \mid \lambda B(\mathcal{X}_{n+1}) \subset K\},$$

   where $\mathcal{X}_{n+1} \subset \mathcal{X}$ is any $(n+1)$ dimensional subspace of $\mathcal{X}$, $B(\mathcal{X}_{n+1})$ is the unit ball in $\mathcal{X}_{n+1}$ (metric in $\mathcal{X}$).

2. Let $T : K \to \mathbb{R}^n$ that $T(f) = \mathcal{E}_n(f) - \mathcal{E}_n(-f)$, then T is odd on $\lambda\partial B(\mathcal{X}_{d+1})$, which means there exists $f^* \in \partial\lambda B(\mathcal{X}_{n+1})$ that $T(f^*) = 0$ because of the famous *Borsuk-Ulam* theorem. Thus,

$$2f^* = (f^* - \mathcal{D}_n(\mathcal{E}_n(f^*))) - (-f^* - \mathcal{D}_n(\mathcal{E}_n(-f^*))).$$

   Hence (by pigeonhole principle), there exists $f^\dagger = f^*$ (or $-f^*$) that

$$\|f^\dagger - \mathcal{D}_n(\mathcal{E}_n(f^\dagger))\|_{\mathcal{X}} \geq \lambda = b_n(K;\mathcal{X}).$$

This holds for any $\mathcal{E}_n$ and $\mathcal{D}_n$, thus $h_n(K;\mathcal{X}) \geq b_n(K;\mathcal{X})$.

3. We see that $b_n(K;\mathcal{X})$ is independent of $\mathcal{E}_n$ and $\mathcal{D}_n$. The proof is standard by construction.

   Let $K = \{f \in W^{m,p}(X) \mid \|f\|_{W^{m,p}} \leq 1\}$, which is compact in $\mathcal{X} = L^p$. Take ϕ as a $C_c^\infty[0,1]^d$ function (e.g., standard bump function). Decompose the domain $X \subset \mathbb{R}^d$ is into disjoint small cells of size ℓ and shift/rescale ϕ to each cell (denoted by ϕ_i^*), then it automatically defines $n = \mathcal{O}(\ell^{-d})$ dimensional subspace of $W^{m,p}(X)$ through

$$\mathcal{X}_{n+1} = \left\{ \sum_i c_i \phi_i^* \mid c_i \in \mathbb{R} \right\}.$$

   For any $f \in \partial B(\mathcal{X}_{n+1})$, it satisfies that

$$\|f\|_{L^p}^p = \left(\sum_i |c_i|^p \right) \|\phi\|_{L^p}^p = 1.$$

   and we need to find the largest λ that

$$\|\lambda f\|_{W^{m,p}}^p = \lambda^p \sum_{|\alpha| \leq m} \left(\ell^{-p|\alpha|} \sum_i |c_i|^p \right) \|\partial^\alpha \phi\|_{L^p}^p \leq 1.$$

   Therefore, $\lambda \asymp \ell^m \asymp n^{-m/d}$.

□

Corollary 1.1.50. *Let $\mathcal{A}_n$ be an algorithm (or map) to compute the parameterization θ of a shallow neural network of width n for an input function $f \in L^2(X)$. If $\mathcal{A}$ is **continuous**, then there exists $f \in W^{m,2}(X)$ that $\|f\|_{W^{m,2}} \leq 1$ and*

$$\|f - \mathcal{A}_n f\|_{L^2} \geq C n^{-m/d}.$$

Remark 1.1.51. *Therefore, the optimal error bound $\mathcal{O}(n^{-m/(d-1)})$ that appears in n-term shallow neural network (see Corollary 1.1.47) cannot be implemented by a **continuous** algorithm.*

Remark 1.1.52. ***Iterative** algorithms do not mean they are continuous.*

Error Bound for General Activations

The following theorem implies that a general activation function can achieve the (stable approximation) error of $\mathcal{O}(n^{-m/d})$ for Sobolev spaces.

Theorem 1.1.53 (Mhaskar (1996)). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ and $\sigma \in C^\infty(I)$ on some interval I but not a polynomial on I.*

$$\mathcal{S} := \left\{ \sum_{i=1}^{n} c_i \sigma(w_i \cdot x + b_i) \mid c_i \in \mathbb{R},\ w_i \in \mathbb{R}^d,\ b_i \in \mathbb{R} \right\}.$$

Then, for any $f \in W^{m,2}(X)$ with bound $\|f\|_{W^{m,2}(X)} \leq 1$,

$$\inf_{g \in \mathcal{S}} \|f - g\|_{L^2(X)} \leq C n^{-m/d}.$$

Rough Idea. Using Theorem 1.1.37 (and its proof in Section 1.1.3), the set

$$\bigcup_{l=0}^{k} \mathcal{P}_l = \left\{ \sum_{i=1}^{s} k_i(w_i \cdot x) \mid k_i \in \bigcup_{l=0}^{k} \pi_l \right\}$$

provides an approximation error of $\mathcal{O}(k^{-m})$, where $s \approx \mathcal{O}(k^{d-1})$. Each $k_i(w_i \cdot x)$ has $\deg(k_i) \leq k$, and can be represented by a limiting finite difference scheme of $k+1$ terms (why?)

$$k_i(w_i \cdot x) = \lim_{\varepsilon \to 0} \sum_{j=1}^{k+1} c_{i,j,\varepsilon} \sigma(w_i \cdot x + b_{i,j,\varepsilon}).$$

Therefore,

$$\bigcup_{l=0}^{k} \mathcal{P}_l \subset \overline{\left\{ \sum_{i=1}^{s(k+1)} c_i \sigma(w_i \cdot x + b_i) \;\middle|\; c_i, b_i \in \mathbb{R}, w_i \cdot \mathbb{R}^d \right\}}.$$

Therefore, the total number of terms is $\mathcal{O}(k^d)$, meaning the approximation error is $\mathcal{O}(n^{-m/d})$. □

Question 1.1.54. *Complete this proof.*

Remark 1.1.55. *This theorem shows that a shallow neural network can first approxi-*

mate the polynomial basis in $\mathcal{P}_k$ and then combine the basis elements to achieve $\mathcal{O}(n^{-m/d})$ error. Note that this process is entirely linear and thus depends continuously on the function to be approximated.

1.1.4 ReLU Bounds

We now consider specific activation functions within the ReLU class. The ReLU (rectified linear unit) is a common activation function defined by

$$\sigma(x) = \max(0, x). \tag{1.4}$$

More generally, we denote $\sigma^k(x) = (\max(0, x))^k$, then $k = 0$ means the Heaviside function. Since the ReLU class is affine, we only consider $w \in \mathbb{S}^{d-1}$ and $|b| \leq 2\operatorname{diam}(X)$. Notice that the ReLU class does not satisfy the requirements in Theorem 1.1.53.

Remark 1.1.56. *In 1D, the ReLU class is precisely the set of spline functions. The results are classical from finite elements. Let*

$$\mathcal{S} = \left\{ \sum_{i=1}^{n} c_i \sigma^k(x + b_i)) \;\middle|\; b_i, c_i \in \mathbb{R} \right\},$$

then for $1 \leq m \leq k + 1$,

$$\inf_{g \in \mathcal{S}} \|f - g\|_{L^2} \leq C n^{-m} \|f\|_{W^{m,2}}.$$

This is different from Theorem 1.1.53, which allows all m. Since $g^{(m)} \equiv 0$ for higher $m \geq k + 1$, which means $g \in \mathcal{S}$ cannot affect higher derivatives.

In Petrushev (1998), it shows that the ridge approximation error in $\mathbb{R}^1$ can imply the neural network's approximation error in $\mathbb{R}^d$.

Theorem 1.1.57 (Petrushev (1998)). *Suppose $1 \leq m \leq \frac{2k+d+1}{2}$, and $f \in W^{m,2}(X)$ and $\|f\|_{W^{m,2}(X)} \leq 1$. Let*

$$\mathcal{S} = \left\{ \sum_{i=1}^{n} c_i \sigma^k(w_i \cdot x + b_i)) \;\middle|\; w_i \in \mathbb{S}^{d-1}, b_i, c_i \in \mathbb{R} \right\}.$$

Then

$$\inf_{g \in \mathcal{S}} \|f - g\|_{L^2(X)} \leq C n^{-m/d}.$$

Remark 1.1.58. *In addition to the classical results in Sobolev spaces, we can consider using σ^k as the activation function, and $f(x) = \mathbb{E}_{(w,b)\sim\Theta}[\sigma^k(w \cdot x + b)]$, as discussed at the beginning of this chapter. The result for $k = 0$ is known by Theorem 1.1.22 and Theorem 1.1.26, the L^2 approximation bound is $\mathcal{O}(n^{-\frac{1}{2}-\frac{1}{2d}})$.*

In conclusion, shallow neural networks already suffer from the **curse of dimensionality** due to the nature of approximation error, making it impractical to approximate general functions in high dimensions with them. In the next section, we will show that multi-layer neural networks can overcome the limitations of representation (without defeating the curse of dimensionality).

1.2 Multi-Layer Perceptron

We first focus on the two-hidden-layer MLP model. The main difference between MLPs and shallow neural networks is that the additional layer provides substantially greater flexibility for approximation when we do not impose strict requirements on the activation function.

Indeed, the shallow neural network structure intrinsically has difficulty representing locally supported functions in $\mathbb{R}^d$, $d \geq 2$. However, with an additional activation layer, the MLP can approximate locally supported functions.

Question 1.2.1. *Try to construct a function*

$$\sigma_h\left(\sum_{i=1}^{n} c_i \sigma_h(w_i \cdot x + b_i)\right)$$

to be compactly supported in the unit cube, where σ_h is the Heaviside function.

1.2.1 Kolmogrov Superposition Theorem

The KST yields a negative answer to Hilbert's 13th problem, which is one of the 23 Hilbert problems compiled in 1900 by David Hilbert.

Problem 1.2.2 (Hilbert's 13th Problem)**.** *Prove whether a solution exists for all 7th-degree equations using continuous functions of two arguments.*

The KST theorem is also called the *Kolmogorov-Arnold Representation theorem*, which states that every multivariate continuous function $f : [0, 1]^n \to \mathbb{R}$ can be **represented** as a superposition of continuous single-variable functions.

Theorem 1.2.3 (KST, Original). *For any continuous function $f : [0,1]^d \to \mathbb{R}$, there exists continuous functions $\phi_{ij} : [0,1] \to \mathbb{R}$, and $\Psi_j \in C[0,1]$, $i \in [n]$, $j \in [2d+1]$, such that*

$$f(x_1, \cdots, x_d) = \sum_{j=1}^{2d+1} \Psi_j \left(\sum_{i=1}^{d} \phi_{ij}(x_i) \right). \tag{1.5}$$

The functions ϕ_{ij} are universal (independent of f).

It can be shown that the outer functions Ψ_j can be shared. The inner functions can also be shared if shifts are allowed. The inner functions can be made Lipschitz continuous, but the outer function cannot be more regular.

Theorem 1.2.4 (KST, Variant). *For any continuous function $f : [0,1]^d \to \mathbb{R}$, there exists constants c_i, $i \in [d]$, and strictly increasing continuous functions $\phi_j : [0,1] \to [0,1]$, $j \in [2d+1]$, and $\Psi \in C[0,1]$ such that*

$$f(x_1, \cdots, x_d) = \sum_{j=1}^{2d+1} \Psi \left(\sum_{i=1}^{d} c_i \phi_j(x_i) \right). \tag{1.6}$$

Rough Idea. We list only a few key steps.

1. Given some $\lambda \in (0,1)$. Let U_f be the set of open tubes $(\phi_j)_{j\in[2d+1]} \subset C([0,1])^{\otimes(2d+1)}$ such that there exists some continuous g with uniform bound $\|g\| \leq \mu\|f\|$, $\mu \in (0,1)$ and

$$\left\| f - \sum_{j=1}^{2d+1} g \left(\sum_{i=1}^{d} c_i \phi_j(x_i) \right) \right\| \leq \lambda \|f\|.$$

   The set U_f is an open, dense subset for any continuous f.

2. Choose a (dense) sequence of continuous functions h_k, and the corresponding sets U_{h_k}. Then, $\bigcap_k U_{h_k}$ is an open, dense subset. One can choose a specific set of tubes $(\phi_j^*)_{j\in[2d+1]} \subset C([0,1])^{\otimes(2d+1)}$. Then for any continuous

f, there exists h_m arbitrarily close, and corresponding continuous g_m,

$$\begin{aligned}&\left\|f-\sum_{j=1}^{2d+1}g\left(\sum_{i=1}^{d}c_i\phi_j^*(x_i)\right)\right\|\\&\leq\|f-h_m\|+\left\|h_m-\sum_{j=1}^{2d+1}g_m\left(\sum_{i=1}^{d}c_i\phi_j^*(x_i)\right)\right\|\\&\leq(\varepsilon+\lambda)\|h_m\|\leq\frac{(\varepsilon+\lambda)}{1-\varepsilon}\|f\|.\end{aligned}$$

3. Finally, take such a g function iteratively, and construct a sequence that converges to f.

□

Remark 1.2.5. *There are various versions of KST, but the representation size of $2d+1$ terms cannot be improved.*

1.2.2 Two-Hidden Layers

The MLP model allows the neural network to approximate a continuous function with arbitrary accuracy if the activation function is slightly *pathological*.

Theorem 1.2.6 (Maiorov and Pinkus (1999)). *There exists an activation function $\sigma\in C^\infty$, strictly increasing, sigmoidal, and for any $f\in C[0,1]^d$ that*

$$\inf_\theta\sup_{x\in[0,1]^d}\left|f(x)-\sum_{i=1}^{4d+3}W_i\sigma\left(\sum_{j=1}^{2d+1}c_{ij}\sigma\left(w_{ij}\cdot x+b_{ij}\right)+B_i\right)\right|=0.$$

Rough Idea. This theorem leverages the pathological activation function to rewrite the KST. Note that the number of inner/outer functions might be improved.

The continuous function Ψ can be approximated by (see Theorem 1.1.45)

$$\Psi(t)\approx a\sigma(t+s_\Psi)+b\sigma(t+s_0)+c\sigma(t+s_1),$$

where σ is the pathological activation function, the shifts s_Ψ depends on Ψ, the other two shifts s_0,s_1 are fixed for linear functions.

Then, replace t with $\sum_i c_i\phi_j(x_i)$ for each $j\in[2d+1]$, we can obtain $2d+1$ functions (depends on Ψ), $2d+1$ linear functions (shift s_0), and $2d+1$ linear

functions (shift s_1). The $2d+1$ linear functions can be merged into $d+1$ linear functions by a linear basis. Thus, in total we need $2d+1+2(d+1)=4d+3$ outer functions.

Similar ideas apply to the inner functions. □

Remark 1.2.7. *We can make $\{w_{ij}\}_{i\in[2d+1]}$ explicit. There are d of them as standard base vectors and $d+1$ of them as unit vectors (for a linear basis)*

Remark 1.2.8. *Of course, such an approximation cannot be stable, which means finding the parameterization θ could be very difficult.*

Extended Reading

Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133.

Canuto, C. and Quarteroni, A. (1982). Approximation results for orthogonal polynomials in sobolev spaces. *Mathematics of Computation*, 38(157):67–86.

DeVore, R. A., Howard, R., and Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478.

DLMF (2025). *NIST Digital Library of Mathematical Functions*. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Guo, B.-y., Shen, J., and Xu, C.-l. (2005). Generalized laguerre approximation and its applications to exterior problems. *Journal of computational Mathematics*, pages 113–130.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

Jones, L. K. (1992). A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The annals of Statistics*, pages 608–613.

Maiorov, V. and Pinkus, A. (1999). Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25(1-3):81–91.

Maiorov, V. E. (1999). On best approximation by ridge functions. *Journal of Approximation Theory*, 99(1):68–94.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory*, 85(1):98–109.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1):164–177.

Petrushev, P. P. (1998). Approximation by ridge functions and neural networks. *SIAM Journal on Mathematical Analysis*, 30(1):155–189.

Siegel, J. W. and Xu, J. (2024). Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537.