Approximation theory in neural networks

Yanhui Su[†]
yanhui_su@brown.edu

March 30, 2018

Outline

Outline

Outline

Outline

Outline

Outline

Basic Notations

1. d: dimension of input layer;
2. L: number of layers;
3. N_l: number of neuros in the lth layers, $l = 1, \cdots, L$;
4. $\rho : \mathbb{R} \to \mathbb{R}$: activation function;
5. $W_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}, 1 \leq l \leq L,\ x \to A_l x + b_l$;
6. $(A_l)_{ij}, (b_l)_i$: the networks weights;

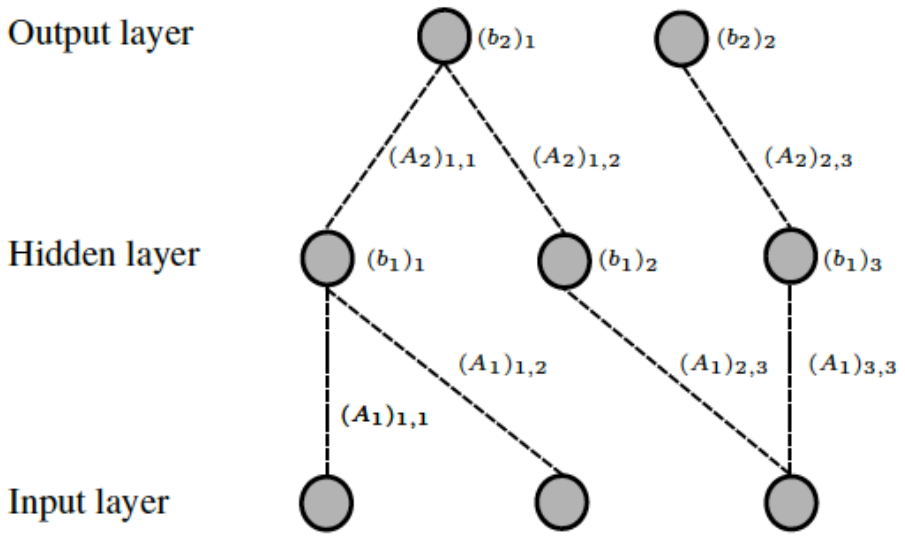
Definition 1

A map $\Phi : \mathbb{R}^d \to \mathbb{R}^L$ given by

$$\Phi(x) = W_L \rho(W_{L-1} \rho(\cdots \rho(W_1(x)))), \qquad x \in \mathbb{R}^d,$$

is called a **neural network**.

Basic Notations



$$A_2 = \begin{pmatrix} (A_2)_{1,1} & (A_2)_{1,2} & 0 \\ 0 & 0 & (A_2)_{2,3} \end{pmatrix}$$

$$A_1 = \begin{pmatrix} (A_1)_{1,1} & (A_1)_{1,2} & 0 \\ 0 & 0 & (A_1)_{2,3} \\ 0 & 0 & (A_1)_{3,3} \end{pmatrix}$$

A classical result of Cybenko

We say the σ is sigmoidal if $\sigma(x) \to \begin{cases} 1, & x \to +\infty \\ 0, & x \to -\infty \end{cases}$.

A classical result on approximation of neural networks is:

Theorem 2

(Cybenko [6]) Let σ be any continuous sigmoidal function. Then finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j \cdot x + \theta_j) \tag{1}$$

are dense in $C(I_d)$.

In [5], T.P. Chen, H. Chen and R.W. Liu gave a constructive proof which only assume that σ is bounded sigmoidal function.

Outline

Approximations of continuous functionals on L^p space

Theorem 3

(Chen and Chen [3]) Suppose that U is a compact set in $L^p[a,b](1<p<\infty)$, f is a continuous functional defined on U, and $\sigma(x)$ is a bounded sigmoidal function, then for any $\varepsilon>0$, there exist $h>0$, a positive integer m, $m+1$ points $a=x_0<x_1<\cdots<x_m=b$, $x_j=a+j(b-a)/m$, $j=0,1,\cdots,m$, a positive integer N and constants c_i, θ_i, $\xi_{i,j}$, $i=1,\cdots,N$, $j=0,1,\cdots,m$, such that

$$\left| f(u) - \sum_{i=1}^{N} c_i \sigma \left(\sum_{j=0}^{m} \xi_{i,j} \frac{1}{2h} \int_{x_j-h}^{x_j+h} u(t)dt + \theta_i \right) \right| < \varepsilon$$

holds for all $u \in U$. Here it is assumed that $u(x)=0$, if $x \notin [a,b]$.

Approximations of continuous functionals on $C[a,b]$

Theorem 4

(Chen and Chen [3]) Suppose that U is a compact set in $C[a,b]$, f is a continuous functional defined on U, and $\sigma(x)$ is a bounded sigmoidal function, then for any $\varepsilon > 0$, there exist $m+1$ points $a = x_0 < \cdots < x_m = b$, a positive integer N and constants c_i, θ_i, $\xi_{i,j}$, $i = 1, \cdots, N$, $j = 0, 1, \cdots, m$, such that for any $u \in U$,

$$\left| f(u) - \sum_{i=1}^{N} c_i \sigma \left(\sum_{j=0}^{m} \xi_{i,j} u(x_j) + \theta_i \right) \right| < \varepsilon$$

An example in dynamical system

Suppose that the input $u(x)$ and the output $s(x) = G(u(x))$ satisfies

$$\frac{ds(x)}{dx} = g(s(x), u(x), x), \quad s(a) = s_0$$

where g satisfies Lipschitz condition, then

$$(Gu)(x) = s_0 + \int_a^x g((Gu)(t), u(t), t)dt.$$

It can be shown that G is a continuous functional on $C[a, b]$. If the input set $U \subset C[a, b]$ is compact, then the output at a specified time d can be approximated by

$$\sum_{i=1}^{N} c_i \sigma \left(\sum_{j=1}^{m} \xi_{i,j} u(x_i) + \theta_i \right).$$

Outline

Approximation by Arbitrary Functions

Definition 5

If a function $g : \mathbb{R} \to \mathbb{R}$ satisfies that all the linear combinations of the form

$$\sum_{i=1}^{N} c_i g(\lambda_i x + \theta_i), \quad \lambda_i, \theta_i, c_i \in \mathbb{R}, i = 1, \cdots, N$$

are dense in every $C[a, b]$, then g is called a **Tauber-Wiener** (TW) function.

Theorem 6

(Chen and Chen [4]) Suppose that $g(x) \in C(\mathbb{R}) \cap \mathcal{S}'(\mathbb{R})$, then $g \in (\mathrm{TW})$ if and only if g is not a polynomial.

Approximation by Arbitrary Functions

Theorem 7

(Chen and Chen [4]) Suppose that K is a compact set in $\mathbb{R}^d$, U is a compact set in $C(K)$, $g \in (\mathrm{TW})$, then for any $\varepsilon > 0$, there are a positive integer N, $\theta_i \in \mathbb{R}$, $\omega_i \in \mathbb{R}^d$, $i = 1, \cdots, N$, which are all **independent** *of $f \in U$ and constants $c_i(f)$ depending on f, $i = 1, \cdots, N$, such that*

$$\left| f(x) - \sum_{i=1}^{N} c_i(f) g(\omega_i \cdot x + \theta_i) \right| < \varepsilon$$

holds for all $x \in K$, $f \in U$. Moreover, every $c_i(f)$ is a **continuous functional** *defined on U.*

Approximation to functionals by Arbitrary Functions

The following theorem can be viewed as a generalization of Theorem 4 of sigmoidal function case.

Theorem 8

(Chen and Chen [4]) Suppose that $g \in (\mathrm{TW})$, X *is a Banach space,* $K \subset X$ *is a compact set,* V *is a compact set in* $C(K)$, f *is a continuous functional defined on* V. *Then for any* $\varepsilon > 0$, *there are positive integers* N, m *points* $x_1, \cdots, x_k \in K$, *and constants* $c_i, \theta_i, \xi_{ij} \in \mathbb{R}$, $i = 1, \cdots, N$, $j = 1, \cdots, m$, *such that*

$$\left| f(u) - \sum_{i=1}^{N} c_i g\left(\sum_{j=1}^{m} \xi_{ij} u(x_j) + \theta_i \right) \right| < \varepsilon$$

holds for all $u \in V$.

Approximation to operators by Arbitrary Functions

Theorem 9

(Chen and Chen [4]) Suppose that $g \in (\mathrm{TW})$, X *is a Banach space,* $K_1 \subset X$, $K_2 \subset \mathbb{R}^d$ *are two compact sets.* V *is a compact set in* $C(K_1)$, G *is a nonlinear continuous operators, which maps* V *to* $C(K_2)$. *Then for any* $\varepsilon > 0$, *there are a positive integers* M, N, m, *constants* $c_i^k, \zeta_k, \xi_{ij}^k \in \mathbb{R}$, *points* $\omega_k \in \mathbb{R}^d$, $x_j \in K_1$, $i = 1, \cdots, M$, $k = 1, \cdots, N$, $j = 1, \cdots, m$, *such that*

$$\left| G(u)(y) - \sum_{i=1}^{M} \sum_{k=1}^{N} c_i^k g\left(\sum_{j=1}^{m} \xi_{ij}^k u(x_j) + \theta_i^k \right) \cdot g(\omega_k \cdot y + \zeta_k) \right| < \varepsilon$$

holds for all $u \in V$, $y \in K_2$.

Outline

Basic notations

1. $\tilde{F}(d\omega) = e^{i\theta(\omega)}F(d\omega)$: the Fourier distribution (i.e. complex-valued measure) of a function $f(x)$ on $\mathbb{R}^d$, and

$$f(x) = \int e^{i\omega\cdot x}\tilde{F}(d\omega) \tag{2}$$

2. B: bounded set in $\mathbb{R}^d$ that contains $\{0\}$
3. Γ_B: the set of functions f on B for which the representation (2) holds for $x \in B$ for some complex-valued measure $\tilde{F}(d\omega)$ for which $\int |\omega| F(d\omega)$ is finite.
4. $\Gamma_{C,B}$: the set of all functions f in Γ_B such that for some $\tilde{F}$ representing f on B

$$\int |\omega|_B F(d\omega) \leq C$$

where $|\omega|_B = \sup_{x\in B} |\omega \cdot x|$.

Universal approximation bounds

Theorem 10

(Barron [1]) For every function f in $\Gamma_{C,B}$, every sigmoidal function σ, every probability measure μ, and every $n \geq 1$, there exists a linear combination of sigmoidal functions $f_n(x)$, such that

$$\int_B (\bar{f}(x) - f_n(x))^2 \mu(dx) \leq \frac{(2C)^2}{n}$$

where $\bar{f}(x) = f(x) - f(0)$.

In theorem 10, the approximation result was proved without the restrictions on $|y_j|$ which yield a difficult problem of searching an unbounded domain.

Universal approximation bounds

Given $\tau > 0$, $C > 0$ and a bounded set B, let

$$G_{\sigma,\tau} = \{\gamma\sigma(\tau(\alpha \cdot x + b)) : |\gamma| \leq 2C, |\alpha|_B \leq 1, |b| \leq 1\}$$

Theorem 11

(Barron [1]) For every $f \in \Gamma_{C,B}$,$\tau > 0$, $n \geq 1$, every probability measure μ, and every sigmoidal function σ with $0 \leq \sigma \leq 1$, there is a function f_n in the convex hull of n functions in $G_{\sigma,\tau}$ such that

$$\left\|\bar{f} - f_n\right\| \leq 2C\left(\frac{1}{n^{1/2}} + \delta_\tau\right)$$

where $\|\cdot\|$ denote the $L^2(\mu, B)$ norm, $\bar{f} = f(x) - f(0)$, and

$$\delta_\tau = \inf_{0<\varepsilon\leq 1/2}\left\{2\varepsilon + \sup_{|z|\geq\varepsilon}|\sigma(\tau z) - 1_{\{z>0\}}|\right\}$$

Outline

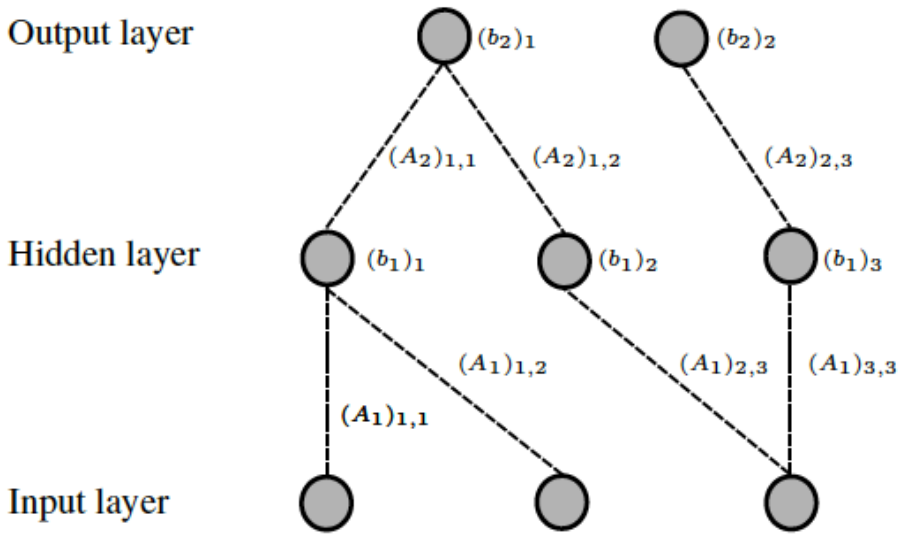
Basic Notations

1. Ω: a domain in $\mathbb{R}^d$;
2. $\mathcal{C}$: a function class in $L^2(\Omega)$.
3. M: the networks connectivity (i.e. the total number of nonzero edge weights). If M is small relative to the number of connections possible, we say that the network is **sparsely connected**.
4. $\mathcal{NN}_{L,M,d,\rho}$: the class of networks $\Phi : \mathbb{R}^d \to \mathbb{R}$ with L layers, connectivity no more than M, and activation function ρ. Moreover, we let

$$\mathcal{NN}_{\infty,M,d,\rho} := \bigcup_{L\in\mathbb{N}} \mathcal{NN}_{L,M,d,\rho},\ \mathcal{NN}_{L,\infty,d,\rho} := \bigcup_{M\in\mathbb{N}} \mathcal{NN}_{L,M,d,\rho}$$

$$\mathcal{NN}_{\infty,\infty,d,\rho} := \bigcup_{M\in\mathbb{N}} \mathcal{NN}_{L,\infty,d,\rho}$$

Basic Notations



$$A_2 = \begin{pmatrix} (A_2)_{1,1} & (A_2)_{1,2} & 0 \\ 0 & 0 & (A_2)_{2,3} \end{pmatrix}$$

$$A_1 = \begin{pmatrix} (A_1)_{1,1} & (A_1)_{1,2} & 0 \\ 0 & 0 & (A_1)_{2,3} \\ 0 & 0 & (A_1)_{3,3} \end{pmatrix}$$

Best M-term Approximation Error

Definition 12

(DeVore and Lorentz, [7]) Given $\mathcal{C} \subset L^2(\Omega)$, and a representation system $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,

$$\Gamma_M^{\mathcal{D}}(f) := \inf_{\substack{I_M \subset \mathbb{N} \\ \sharp I_M = M}} \left\| f - \sum_{i\in I_M} c_i \varphi_i \right\|_{L^2(\Omega)}$$

We call $\Gamma_M^{\mathcal{D}}(f)$ the **best M-term approximation error** of f with respect to $\mathcal{D}$.
The supremal $\gamma > 0$ such that there exists $C > 0$ with

$$\sup_{f\in\mathcal{C}} \Gamma_M^{\mathcal{C}}(f) \leq CM^{-\gamma}, \qquad \forall M \in \mathbb{N}$$

will be referred to as $\gamma^*(\mathcal{C}, \mathcal{D})$.

Best M-term Approximation Error

1. It is conceivable that the optimal approximation rate for $\mathcal{C}$ in any representation system reflects specific properties of $\mathcal{C}$. However, a countable and dense repersentation system $\mathcal{D} \subset L^2(\mathbb{R}^d)$ results in $\gamma^*(\mathcal{C}, \mathcal{D}) = \infty$.
2. In numerical computation, we need to find some efficient methods to approximate any $f \in \mathcal{C}$ by linear combination of finite elements in $\mathcal{D}$. However, finding index in the full index $\mathbb{N}$ is computationally infeasible.
3. In [8], Donoho suggests to restrict the search for the optimal coefficient set to the first $\pi(M)$ coefficients where π is some polynomial. This approach is known as **polynomial-depth search**.

Effective Best M-term Approximation Error

To overcome these problems, Donoho [8] and Grohs [9] proposed the following

Definition 13

Given $\mathcal{C} \subset L^2(\Omega)$, a representation system $\mathcal{D} = (\phi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$. For $\gamma > 0$, we say that $\mathcal{C}$ has **effective best M-term approximation rate** $M^{-\gamma}$ in $\mathcal{D}$ if there exists a univariate polynomial π and constants $C, D > 0$ such that for all $M \in \mathbb{N}$ and $f \in \mathcal{C}$,

$$\left\| f - \sum_{i\in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \leq C M^{-\gamma}$$

for some index set $I_M \subset \{1, \cdots, \pi(M)\}$ with $\sharp I_M = M$ and cofficients $(c_i)_{i\in I_M}$ satisfying $\max_{i\in I_M} |c_i| \leq D$. The supremal $\gamma > 0$ such that $\mathcal{C}$ has effective best M-term approximation rate $M^{-\gamma}$ in $\mathcal{D}$ will be referred to as $\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$.

Best M-edge Approximation Error

Definition 14

(Bölcskei et. al. [2]) Given $\mathcal{C} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,

$$\Gamma_M^{\mathcal{NN}}(f) := \inf_{\Phi \in \mathcal{NN}_{\infty,M,d,\rho}} \|f - \Phi\|_{L^2(\Omega)}$$

We call $\Gamma_M^{\mathcal{NN}}(f)$ the **best M-edge approximation error** of f. The supremal $\gamma > 0$ such that a $C > 0$ with

$$\sup_{f \in \mathcal{C}} \Gamma_M^{\mathcal{NN}}(f) \leq C M^{-\gamma}, \qquad \forall M \in \mathbb{N}$$

will be referred to as $\gamma^*_{\mathcal{NN}}(\mathcal{C}, \rho)$.

Best M-edge Approximation Error

The following theorem in [10] shows that Definition 14 has the samilar troubles with the Definition 12.

Theorem 15

There exists a function $\rho : \mathbb{R} \to \mathbb{R}$ that is C^∞, strictly increasing, and satisfies $\lim_{x\to\infty} \rho(x) = 1$ and $\lim_{x\to-\infty} \rho(x) = 0$, such that for any $d \in \mathbb{N}$, any $f \in C([0,1]^d)$ and any $\varepsilon > 0$ there exists a neural network Φ with activation function ρ three layers of dimensions $N_1 = 3d$, $N_2 = 6d + 3$, and $N_3 = 1$ satisfying

$$\sup_{x\in[0,1]^d} |f(x) - \Phi(x)| \leq \varepsilon$$

Effective Best M-edge Approximation Error

Definition 16

(Bölcskei et. al. [2]) For $\gamma > 0$, $\mathcal{C} \subset L^2(\Omega)$ is said to have **effective best M-edge approximation rate $M^{-\gamma}$** by neural networks with activation function ρ if there exist $L \in \mathbb{N}$, a univariate polynomial π, and a constant $C > 0$ such that for all $M \in \mathbb{N}$ and $f \in \mathcal{C}$

$$\|f - \Phi\|_{L^2(\Omega)} \leq CM^{-\gamma}$$

for some $\Phi \in \mathcal{NN}_{L,M,d,\rho}$ with the weights of Φ all bounded in absolute value by $\pi(M)$.
The supremal $\gamma > 0$ such that $\mathcal{C}$ has effective best M-edge approximation rate $M^{-\gamma}$ will henceforth be denoted as $\gamma^{*,\text{eff}}_{\mathcal{NN}}(\mathcal{C}, \rho)$.

Min-Max Rate Distortion Theory in [8, 9]

Definition 17

Let $\mathcal{C} \subset L^2(\Omega)$, for each $l \in \mathbb{N}$, we denote by

$$\mathfrak{E}^l := \{E : \mathcal{C} \to \{0,1\}^l\}$$

the set of binary encoders of $\mathcal{C}$ of length l, and we let

$$\mathfrak{D}^l := \{D : \{0,1\}^l \to L^2(\Omega)\}$$

be the set of binary decoders of length l. An encoder-decoder pair $(E, D) \in \mathfrak{E}^l \times \mathfrak{D}^l$ is said to **achieve distortion** $\varepsilon > 0$ over the function class $\mathcal{C}$, if

$$\sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon$$

Min-Max Rate Distortion Theory in [8, 9]

Definition 18

Let $\mathcal{C} \subset L^2(\Omega)$, for $\varepsilon > 0$ the minimax code length $L(\varepsilon, \mathcal{C})$ is

$$L(\varepsilon, \mathcal{C}) := \min \left\{ l \in \mathbb{N} : \begin{array}{l} \exists (E, D) \in \mathfrak{E}^l \times \mathfrak{D}^l \text{ such that} \\ \sup_{f \in \mathcal{C}} \| D(E(f)) - f \|_{L^2(\Omega)} \leq \varepsilon \end{array} \right\}$$

Moreover, the **optimal exponent** $\gamma^*(\mathcal{C})$ is defined by

$$\gamma^*(\mathcal{C}) := \inf\{\gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) = O(\varepsilon^{-\gamma})\}$$

Min-Max Rate Distortion Theory in [8, 9]

Theorem 19

Let $\mathcal{C} \subset L^2(\Omega)$, and the optimal effective best M-term approximation rate of $\mathcal{C}$ in $\mathcal{D} \subset L^2(\Omega)$ be $M^{-\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})}$. Then,

$$\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) \leq \frac{1}{\gamma^*(\mathcal{C})}.$$

If the representation system $\mathcal{D}$ satifies

$$\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) = \frac{1}{\gamma^*(\mathcal{C})},$$

then, $\mathcal{D}$ is said to be optimal for the function class $\mathcal{C}$.

Foundamental Bound on Effective M-edge Approximation

Theorem 20

(Bölcskei et. al. [2]) Let $\mathcal{C} \subset L^2(\Omega)$ and

$$\mathbf{Learn} : (0,1) \times \mathcal{C} \to \mathcal{NN}_{\infty,\infty,d,\rho}$$

be a map such that, for each pair $(\varepsilon, f) \in (0,1) \times \mathcal{C}$, every weight of the neural network $\mathbf{Learn}(\varepsilon, f)$ can be encoded with no more than $c \log_2(\varepsilon^{-1})$ bits while guaranteeing that

$$\sup_{f \in \mathcal{C}} \|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$$

Then

$$\sup_{\varepsilon \in (0, \frac{1}{2})} \varepsilon^{\frac{1}{\gamma}} \cdot \sup_{f \in \mathcal{C}} \mathcal{M}(\mathbf{Learn}(\varepsilon, f)) = \infty, \qquad \forall \gamma > \frac{1}{\gamma^*(\mathcal{C})}$$

Foundamental Bound on Effective M-edge Approximation

The main idea of the proof of Theorem 20 is encoding the topology and weights of the map $\mathbf{Learn}(\varepsilon, f)$ by encoder-decoder pairs $(E, D) \in \mathfrak{E}^{l(\varepsilon)} \times \mathfrak{D}^{l(\varepsilon)}$ achieving distortion ε over $\mathcal{C}$ with

$$l(\varepsilon) \leq C_0 \cdot \sup_{f \in \mathcal{C}} \mathcal{M}(\mathbf{Learn}(\varepsilon, f)) \log_2(\mathcal{M}(\mathbf{Learn}(\varepsilon, f))) \log_2 \left(\frac{1}{\varepsilon}\right),$$

where $C_0 > 0$ is a constant.

Foundamental Bound on Effective M-edge Approximation

Corollary 21

(Bölcskei et. al. [2]) Let $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{C} \subset L^2(\Omega)$. Then, for all $\rho : \mathbb{R} \to \mathbb{R}$ that are Lipschitz continuous or differentiable with polynomially bounded first derivative, we have

$$\gamma^{*,\text{eff}}_{\mathcal{NN}}(\mathcal{C}, \rho) \leq \frac{1}{\gamma^*(\mathcal{C})}$$

We call a function class $\mathcal{C} \subset L^2(\Omega)$ optimally representable by neural networks with activation function $\rho : \mathbb{R} \to \mathbb{R}$, if

$$\gamma^{*,\text{eff}}_{\mathcal{NN}}(\mathcal{C}, \rho) = \frac{1}{\gamma^*(\mathcal{C})}$$

From Representation Systems to Neural Networks

Definition 22

(Bölcskei et. al. [2]) Let $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be a representation system. Then, $\mathcal{D}$ is said to be **representable by neural networks** (with activation function ρ), if there exists $L, R \in \mathbb{N}$ such that for all $\eta > 0$ and every $i \in \mathbb{N}$ there is a neural network $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ and

$$\|\varphi_i - \Phi_{i,\eta}\|_{L^2(\Omega)} \leq \eta$$

If, in addition, the neural networks $\Phi_{i,\eta} \in \mathcal{NN}_{L,R,d,\rho}$ have weights that are uniformly polynomially bounded in (i, η^{-1}), and if ρ is either Lipschitz-continuous, or differentiable with polynomially bounded derivative, we call the representation system $(\varphi_i)_{i\in\mathbb{N}}$ **effectively representable by neural networks** (with activation function ρ).

From Representation Systems to Neural Networks

Theorem 23

(Bölcskei et. al. [2]) Let $\Omega \subset \mathbb{R}^d$ be bounded, and suppose that $\mathcal{C} \subset L^2(\Omega)$ is effectively representable in the representation system $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$. Suppose that $\mathcal{D}$ is effectively representable by neural networks. Then, for all $\gamma < \gamma^{*,\mathrm{eff}}(\mathcal{C},\mathcal{D})$ there exist constants $c, L > 0$ and and a map

$$\mathbf{Learn} : (0,1) \times L^2(\Omega) \to \mathcal{NN}_{L,\infty,d,\rho}$$

such that for every $f \in \mathcal{C}$ the following statements hold:

1. *there exists $k \in \mathbb{N}$ such that each weight of the network $\mathbf{Learn}(\varepsilon, f)$ is bounded by ε^{-k}.*
2. *the error bound $\|f - \mathbf{Learn}(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon$ holds true, and*
3. *the neural network $\mathbf{Learn}(\varepsilon, f)$ has at most $c\varepsilon^{-1/\gamma}$ edges.*

From Representation Systems to Neural Networks

Specifically, in [2], the authors show that all function classes that are optimally approximated by a general class of representation systems–so-called **affine systems**–can be approximated by deep neural networks with minimal connectivity and memory requirements. Affine systems encompass a wealth of representation systems from applied harmonic analysis such as wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules.

Reference

[1] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information theory, 930–945, 39(3), 1993.

[2] H. Bölcskei, P. Grohs, G. Kutyniok and P. Petersen, *Optimal approximation with sparsely connected deep neural networks*, arXiv:1705.01714, 2017.

[3] T.P. Chen and H. Chen, *Approximations of continuous functionals by neural networks with application to dynamic systems*, IEEE Transactions on Neural Networks, 910–918, 4(6), 1993.

[4] T.P. Chen and H. Chen, *Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems*, IEEE Transactions on Neural Networks, 911–917, 6(4), 1995.

[5] T.P. Chen, H. Chen and R.W. Liu, *A constructive proof and an extension of Cybenkos approximation theorem*, In Computing science and statistics, 163–168, Springer, 1992.

[6] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems, 303–314, 2(4), 1989.

[7] R.A. DeVore and G.G. Lorentz, *Constructive approximation*, Springer Science & Business Media, 1993.

[8] D. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Applied and computational harmonic analysis, 100–115, 1(1), 1993.

[9] P. Grohs, *Optimally sparse data representations*, In Harmonic and Applied Analysis, 199–248, Springer, 2015.

[10] V. Maiorov and A. Pinkus, *Lower bounds for approximation by MLP neural networks*, Neurocomputing, 81–91, 25(1–3), 1999.

Thank You for Your Attention!