APPROXIMATION THEORY OF THE MLP MODEL IN NEURAL NETWORKS

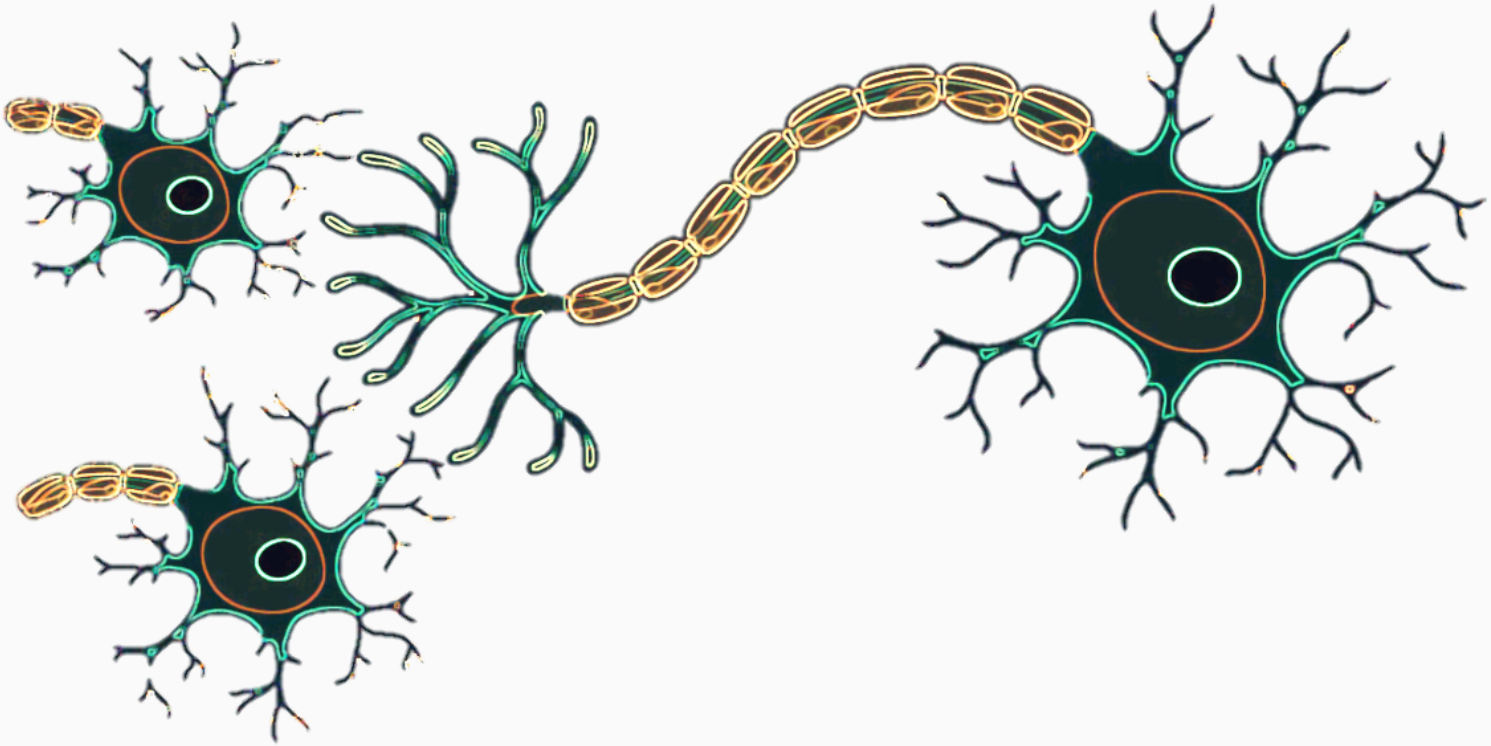
Allan Pinkus

Presented by: Shadi Tasdighi Kalat

November 12, 2019

Introduction

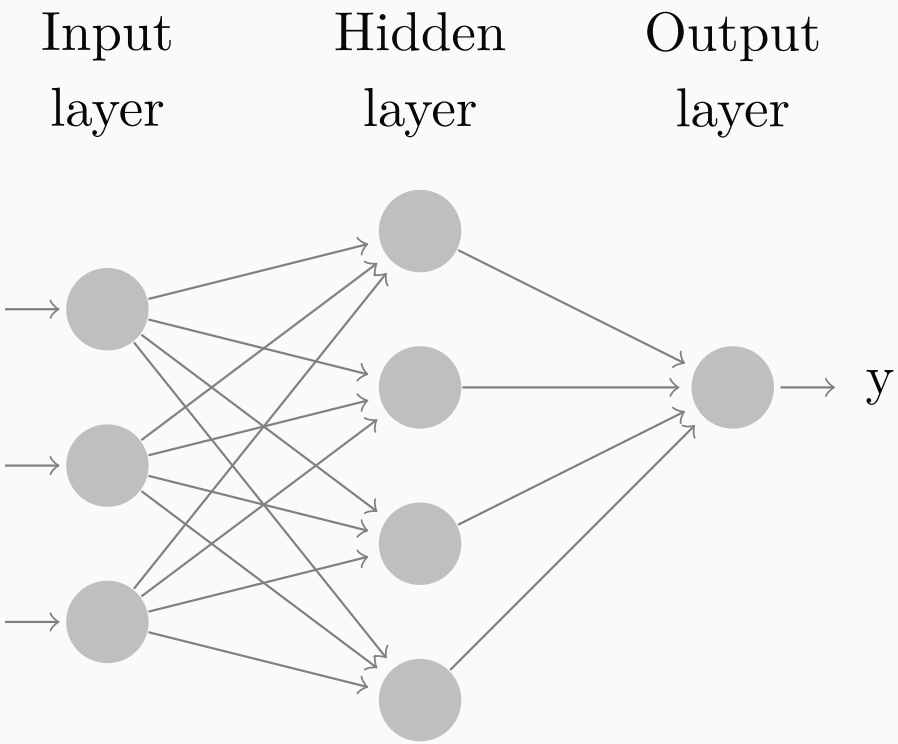
Network of neurons



Introduction

A simplified model

Multilayer feedforward perceptron



History

Perceptron: Neural network with no hidden layer and Heaviside activation function. [1]

With this model two sets of points can be classified if and only if they are linearly separable. To separate N generic points in $\mathbb{R}^n$ by a one-hidden layer model with Heaviside activation, we need at least [N/n] (the smallest integer greater than or equal to N/n) units in the hidden layer.[2]

[1] Rosenblat, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), pp.386-408.
[2] Baum, E.B., 1988. On the capabilities of multilayer perceptrons. Journal of complexity, 4(3), pp.193-215.

Assumptions

1- The output of the jth unit of the input layer is x_{0j}.

$$x_{i+1,k} = \sigma\Big(\sum_j w_{ikj} x_{ij} - \theta_{i,k}\Big). \tag{1}$$

2 - There is no activation function applied to the output layer and there is only one output.

A single hidden-layer

$$y = \sum_{i=1}^{r} c_i \sigma\Big(\sum_{j=1}^{n} w_{i,j} x_j - \theta_i\Big). \tag{2}$$

$w_{i,j}$: weight between the jth unit of the input and the ith unit in the hidden layer.

θ_i is the threshold at the ith unit of the hidden layer.

c_i is the weight between the ith unit of the hidden layer and the output.

Model

For one hidden layer

$$y = \sum_{i=1}^{r} c_i \sigma(w^i.x - \theta_i). \tag{3}$$

For two hidden layers

$$y = \sum_{k=1}^{s} d_k \sigma\Big(\sum_{i=1}^{r} c_{ik} \sigma(w^{ik}.x - \theta_{ik}) - \gamma_k \Big). \tag{4}$$

Density

Consider the set

$$\mathcal{M}(\sigma) = \text{span}\{\sigma(w.x - \theta) : \theta \in \mathbb{R}, w \in \mathbb{R}^n\}. \tag{5}$$

For which σ it is true that $\forall f \in C(\mathbb{R}^n), \text{compact} K \subset \mathbb{R}^n, \epsilon > 0,$

$$\exists g \in \mathcal{M}(\sigma) \quad \text{s.t.} \quad \max_{x \in K} |f(x) - g(x)| < \epsilon \tag{6}$$

Main result

Theorem 3.1
Let $\sigma \in C(\mathbb{R})$. Then $\mathcal{M}(\sigma)$ is dense in $C(\mathbb{R}^n)$, in the topology of uniform convergence on compacta, if and only if σ is not a polynomial.

If σ is a polynomial of degree m, then $\sigma(w.x - \theta)$ is also a polynomial of degree at most m, thus $\mathcal{M}(\sigma)$ is not dense in $C(\mathbb{R}^n)$.

The main part is the converse result.
Is it possible to restrict w and θ, and enlarge the class of eligible σ and still obtain the desired density?

Ridge functions

Definition

Ridge functions are multivariate functions of the form

$$g(a_1 x_1 + \ldots + a_n x_n) = g(x.a) \tag{7}$$

where $g : \mathbb{R} \to \mathbb{R}$ and $a = (a_1, \ldots, a_n) \in \mathbb{R}^n \setminus \{0\}$ is a fixed direction. Set

$$\mathcal{R} = \text{span}\{g(a.x) : a \in \mathbb{R}^n, g : \mathbb{R} \to \mathbb{R}\} \tag{8}$$

$\mathcal{R}$ contains all functions of the form $\cos(a.x)$ and $\sin(a.x)$. So ridge functions have the density property.

Dense subsets of ridge functions include $e^{a.x}$ and $(a.x)^k, \quad k = 0, 1, \ldots$.

Ridge functions

Theorem 3.2

The set of ridge functions

$$\mathcal{R}(A) = \text{span}\{g(a.x) : a \in A, g \in C(\mathbb{R})\} \tag{9}$$

is dense in $C(\mathbb{R}^n)$, if and only if there is no trivial homogeneous polynomial that vanishes on A.

Proposition 3.3

Assume $\Lambda, A \subset \mathbb{R}$ for which

$$\mathcal{N}(\sigma; \Lambda, \Theta) = \text{span}\{\sigma(\lambda t - \theta) : \lambda \in \Lambda, \theta \in \Theta\} \tag{10}$$

is dense in $C(\mathbb{R})$ and A is such that $\mathcal{R}(A)$ is dense in $C(\mathbb{R}^n)$. Then

$$\mathcal{M}(\sigma; \Lambda \times A, \Theta) = \text{span}\{\sigma(w.x - \theta) : w \in \Lambda \times A, \theta \in \Theta\} \tag{11}$$

is dense in $C(\mathbb{R}^n)$

Proof. Let $f \in C(K)$. Since $\mathcal{R}(A)$ is dense in $C(K)$,

$$\forall \epsilon > 0 \quad \exists g_i \in C(\mathbb{R}), \quad a^i \in A, \quad i = 1, \ldots, r \quad \text{s.t.}$$
$$\forall x \in K \quad \left| f(x) - \sum_{i=1}^{r} g_i(a^i.x) \right| < \frac{\epsilon}{2}. \tag{12}$$

Since K is compact, $\{a^i.x : x \in K\} \subseteq [\alpha_i, \beta_i]$. Also, $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C[\alpha_i, \beta_i]$, $\exists c_{ij} \in \mathbb{R}$, $\lambda_{ij} \in \Lambda$, and $\theta_{ij} \in \Theta$, $j = 1, \ldots, m_i$, $i = 1, \ldots, r$ for which

$$\left| g_i(t) - \sum_{j=1}^{m_i} c_{ij} \sigma(\lambda_{i,j} t - \theta_{ij}) \right| < \frac{\epsilon}{2r}, \quad \forall t \in [\alpha_i, \beta_i]$$
$$\left| f(x) - \sum_{i=1}^{r} \sum_{j=1}^{m_i} c_{ij} \sigma(\lambda_{i,j} a^i.x - \theta_{ij}) \right| < \epsilon \quad \forall x \in K. \tag{13}$$

Proposition 3.4
Let $\sigma \in C^\infty(\mathbb{R})$ and assume σ is not a polynomial. Then $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof. Since $\sigma \in C^\infty(\mathbb{R})$ and $[\sigma((\lambda + h)t - \theta_0) - \sigma(\lambda t - \theta_0)]/h \in \mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ for all $h \neq 0$, it follows that:

$$\frac{d^k}{d\lambda^k}\sigma(\lambda t - \theta_0)\Big|_{\lambda=0} = t^k \sigma^{(k)}(-\theta_0) \tag{14}$$

is contained in $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$ for any k. Since $\sigma^{(k)}(-\theta_0) \neq 0$, the set $\overline{\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})}$ contains all polynomials, and by Weierstrass Theorem, $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ in dense in $C(K)$ for every compact $K \subset \mathbb{R}$.

Ridge functions

Corollary 3.5

Let Λ be any set containing a sequence of values tending to zero, and let Θ be any open interval. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be such that $\sigma \in C^\infty(\Theta)$ and not a polynomial on Θ. Then $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C(\mathbb{R})$.

Ridge functions

To weaken the smoothness demands:

Proposition 3.7

Let $\sigma \in C(\mathbb{R})$ and assume σ is not a polynomial. Then $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Now to consider a class of discontinuous functions, same results hold if σ that is bounded and Riemann-integrable on every finite interval.

Ridge functions

Proposition 3.8

Assume $\sigma : \mathbb{R} \to \mathbb{R}$ bounded and Riemann-integrable on every finite interval. Assume σ is not a polynomial. Then $\mathcal{N}(\sigma; \mathbb{R}, \mathbb{R})$ is dense in $C(\mathbb{R})$.

To allow a finite set of gains:

Corollary 3.9

Let Λ be any set containing a sequence of values tending to zero and let Θ be any open interval. Assume $\sigma : \mathbb{R} \to \mathbb{R}$ is bounded and Riemann-integrable on Θ and not a polynomial a.e on Θ. Then $\mathcal{N}(\sigma; \Lambda, \Theta)$ is dense in $C(\mathbb{R})$.

.

Dilations (the set Λ) are not necessary

Proposition 3.10
Assume $\sigma \in C(\mathbb{R}) \cap L^1(\mathbb{R})$ (bounded, continuous and non-decreasing) and not constant. Then $\mathcal{N}(\sigma; 1, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proof.
Assume $\sigma \in C(\mathbb{R}) \cap L^1(\mathbb{R})$. Continuous linear functionals on $C(\mathbb{R}$ are represented by Borel measures of finite total variations and compact support. If $\mathcal{N}(\sigma; 1, \mathbb{R})$ is not dense in $C(\mathbb{R})$, $\exists \mu$ s.t.

$$\int_{-\infty}^{+\infty} \sigma(t-\theta)\, d\mu(t) = 0 \implies \hat{\sigma}(\omega)\hat{\mu}(\omega) = 0, \forall \omega \in \mathbb{R}. \tag{15}$$

$\hat{\mu}$ is an integral function (exponential type) and $\hat{\sigma}$. Therefore $\hat{\sigma}$ must vanish where $\hat{\mu} \neq 0$, this gives $\sigma = 0$.

Mean-periodic functions

Definition

Mean-periodic functions are the solutions of homogeneous convolution equations.[3]

Consider a complex-valued function f of a real variable. The function f is periodic with period a precisely if for all $x \in \mathbb{R}$, we have $f(x) - f(x-a) = 0$. This can be written as:

$$\int f(x-y)\,d\mu(y) = 0 \tag{16}$$

where μ is the difference between the Dirac measures at 0 and a.

A mean-periodic function is a function f for which there exists a compactly supported (signed) Borel measure μ for which $f * \mu = 0$.[4]

[3] Delsarte, J., 1935. Les fonctions moyenne-périodiques. J. Math. Pures Appl, 14(403453), p.9.

[4] Schwartz, L., 1947. Théorie générale des fonctions moyenne-périodiques. Annals of Mathematics, pp.857-929.

Mean-periodic functions

Definition

A function $f \in C(\mathbb{R}^n)$ is mean-periodic if

$$\text{span}\{f(x-a) : a \in \mathbb{R}^n\} \tag{17}$$

is not dense in $C(\mathbb{R}^n)$.

Mean-periodic functions are characterized by the functions of the form $t^m e^{\gamma t}$, where $\gamma \in \mathbb{C}$.

Proposition 3.11

Let $\sigma \in C(\mathbb{R})$, not a polynomial. For any Λ that contains a sequence tending to a finite limit point, the set $\mathcal{N}(\sigma; \Lambda, \mathbb{R})$ is dense in $C(\mathbb{R})$.

Proposition 3.12

Let $\sigma \in C(\mathbb{R})$. If $\sigma \in L^p(\mathbb{R}), 1 \leq p < \infty($ bounded, has a limit at $+\infty$ or $-\infty$ and is not the constant function). Then σ is not mean-periodic.

Interpolation

Assume we are given $\sigma \in C(\mathbb{R})$. For k distinct points $\{x^i\}_{i=1}^k \subset \mathbb{R}^n$, and associated data $\{\alpha^i\}_{i=1}^k \subset \mathbb{R}$, can we always find m, $\{w^j\}_{j=1}^m \subset \mathbb{R}^n$ and $\{c_j\}_{j=1}^m$,$\{\theta_j\}_{j=1}^m \subset \mathbb{R}$ for which

$$\sum_{j=1}^{m} c_j \sigma(w^j.x^i - \theta_j) = \alpha_i, \quad \forall i = 1, \ldots, k? \tag{18}$$

If σ is a sigmoidal, continuous and nondecreasing function, we can always interpolate with $m = k$. [5]

If σ is any bounded, continuous, non-decreasing and nonlinear function which has a limit at $+\infty$ or $-\infty$. [6]

[5] Ito, Y., 1996. Supper position of linearly independent functions and finite mapping by neural networks. Math. Scientists, 21, pp.27-33.

[6] Huang, G.B. and Babri, H.A., 1998. Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions. IEEE transactions on neural networks, 9(1), pp.224-229.

Interpolation

Theorem 5.1 Let $\sigma \in C(\mathbb{R})$ is not a polynomial. For any k distinct points $\{x^i\}_{i=1}^k \subset \mathbb{R}^n$, and associated data $\{\alpha^i\}_{i=1}^k \subset \mathbb{R}$, there exists m, $\{w^j\}_{j=1}^k \subset \mathbb{R}^n$ and $\{c_j\}_{j=1}^k$,$\{\theta_j\}_{j=1}^k \subset \mathbb{R}$ such that

$$\sum_{j=1}^{k} c_j \sigma(w^j.x^i - \theta_j) = \alpha_i, \quad \forall i = 1, \ldots, k. \tag{19}$$

Moreover, if σ is not mean-periodic, then we may choose $\{w^j\}_{j=1}^k \subset S^{n-1}$, where

$$S^{n-1} = \{y : ||y||_2 = 1\}. \tag{20}$$

If σ is a polynomial, then the ability to interpolate depends on the choice of points and the degree of σ.

Degree of approximation

For a given σ, for each r we set

$$\mathcal{M}_r(\sigma) = \Big\{ \sum_{i=1}^{r} c_i \sigma(w^i.x - \theta_i) : c_i, \theta_i \in \mathbb{R}, w^i \in \mathbb{R}^n \Big\}. \tag{21}$$

We know if σ is not a polynomial, then $\forall f \in C(K), \quad \exists g_r \in \mathcal{M}_r(\sigma)$s.t.

$$\lim_{r\to\infty} \max_{x\in K} |f(x) - g_r(x)| = 0. \tag{22}$$

What can we say about the rate of approximation?

Degree of approximation

For functions defined on B^n, the Sobolev space $\mathcal{W}_p^m$ is defined as the completion of $C^m(B^n)$ w.r.t the norm

$$||f||_{m,p} = \begin{cases} (\sum_{0\leq|k|\leq m} ||D^k f||_p^p)^{1/p} & , 1 \leq p < \infty \\ \max_{0\leq|k|\leq m} ||D^k f||_p^p)_\infty & , p = \infty \end{cases} \tag{23}$$

Set

$$\mathcal{B}_p^m := \{f : f \in \mathcal{W}_p^m, ||f||_{m,p} \leq 1\}. \tag{24}$$

Considering the lower bounds:

$$E(f; \mathcal{M}_r(\sigma); X) = \inf_{g\in\mathcal{M}_r(\sigma)} ||f-g||_X \geq \inf_{g\in\mathcal{R}_r(\sigma)} ||f-g||_X = E(f; \mathcal{R}_r(\sigma); X) \tag{25}$$

where

$$\mathcal{R}_r = \Big\{ \sum_{i=1}^{r} g_i(a^i.x) : a^i \in \mathbb{R}^n, g_i \in C(\mathbb{R}), i = 1, \ldots, r. \Big\}. \tag{26}$$

Degree of approximation

Maiorov (1999):[7]
Assume $m \geq 1,\ n \geq 2$. Then $\forall r \quad \exists f \in \mathcal{B}_2^m$ for which

$$Cr^{-m/(n-1)} \leq E(f; \mathcal{R}_r; L_2) \leq Cr^{-m/(n-1)} \tag{27}$$

Theorem 6.2 For each $p \in [1, \infty], m \geq 1, n \geq 2$,

$$E(\mathcal{B}_p^m; \mathcal{R}_r; L_p) \leq Cr^{-m/(n-1)} \tag{28}$$

in which C is independent of r.

Using ridge functions, we can approximate at least as well as we can approximate with any polynomial space contained therein.

[7] Maiorov, V.E., 1999. On best approximation by ridge functions. Journal of Approximation Theory, 99(1), pp.68-94.

Degree of approximation

Proposition 6.3

There exists $\sigma \in C^\infty(\mathbb{R})$, sigmoidal and strictly increasing s.t.

$$\forall g \in \mathcal{R}_r, \epsilon > 0 \quad \exists c_i, \theta_i \in \mathbb{R}, w^i \in \mathbb{R}^n, i = 1, \ldots, r+n+1$$

$$\text{s.t.} \quad \left| g(x) - \sum_{i=1}^{r+n+1} c_i \sigma(w^i.x - \theta_i) \right| < \epsilon \quad \forall x \in B^n. \quad (29)$$

Corollary 6.4

There exists $\sigma \in C^\infty(\mathbb{R})$, sigmoidal and strictly increasing for which

$$E(\mathcal{B}_p^m; \mathcal{M}_r; L_p) \leq C r^{-m/(n-1)}, \quad \forall p \in [1, \infty], m \geq 1, n \geq 2. \quad (30)$$

Degree of approximation

Theorem 6.6

Let $Q_r : L_p \to \mathcal{M}_r(\sigma)$ be any method of approximation where c_i, θ_i and $w^i, \quad i = 1, \ldots, r$ are continuously dependent on f. Then

$$\sup_{f \in \mathcal{B}_p^m} ||f - Q_r f||_p \geq C r^{-m/n}. \tag{31}$$

for some C independent of r.

Theorem 6.7

For logistic sigmoid activation function $\sigma(t) = \dfrac{1}{1 + e^{-t}}$

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \geq C(r \log r)^{-m/n} \tag{32}$$

Upper bounds

If $\sigma \in C^{\infty}(\Theta)$ on some open interval Θ and σ not a polynomial on Θ. Then $\forall p \in [1, \infty], m \geq 1, n \geq 2$

$$E(\mathcal{B}_p^m; \mathcal{M}_r(\sigma); L_p) \leq Cr^{-m/n} \tag{33}$$

Mhaskar (1996)
The optimal order of approximation from $\mathcal{M}_r(\sigma)$ will not be better than what could be obtained by approximating from the polynomial space P_k of dimension $r \asymp k^n$.

Degree of approximation

Petrushev 1998:

For each $k \in \mathbb{Z}_+$, let

$$\sigma_k(t) = \begin{cases} t^k, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{34}$$

Then

$$E(\mathcal{B}_2^m; \mathcal{M}_r(\sigma_k); L_2) \leq Cr^{-m/n} \tag{35}$$

for $m = 1, \ldots, k + 1 + \dfrac{(n-1)}{2}$.

Degree of approximation

Makovoz 1996:
Let K be a bounded subset of a Hilbert space H. Let $f \in \mathrm{co}K$. Then there exists an f_r of the form $f_r = \sum_{i=1}^{r} a_i g_i$ for some $g_i \in K$, $a_i \geq 0, i = 1, \ldots, r$ and $\sum_{i=1}^{r} a_i \leq 1$, satisfying

$$||f - f_r||_H \leq \frac{2\epsilon_r(K)}{\sqrt{r}} \tag{36}$$

where
$\epsilon_r(K) = \inf\{\epsilon > 0 : K \quad \text{can be covered by } r \text{ sets of diameter} \leq \epsilon\}$ for $m = 1, \ldots, k + 1 + \frac{(n-1)}{2}$.
If σ is a piecewise continuous sigmoidal function

$$E(\mathcal{B}_2^{(n+1)/2}; \mathcal{M}_r(\sigma); L_2) \leq Cr^{-(n+1)/2n} \tag{37}$$

Two hidden layers

For a single hidden layer, there is an intrinsic lower bound on the degree of approximation which depends on the number of units used. However, there is no theoretical lower bound on the error of approximation if we allow two hidden layers.

Theorem 7.1

There exist an activation function σ, which is C^∞, strictly increasing, sigmoidal which

$$\forall f \in C[0,1]^n \text{and} \epsilon > 0, \quad \exists d_i, c_{ij}, \theta_{ij}, \gamma_i, w^{i,j} \in \mathbb{R}^n \quad \text{s.t.}$$

$$\left| f(x) - \sum_{i=1}^{4n+3} d_i \sigma\Big(\sum_{j=1}^{2n+1} c_{ij} \sigma(w^{ij}.x + \theta_{ij}) + \gamma_i \Big) \right| < \epsilon, \quad \forall x \in [0,1]^n. \quad (38)$$

Questions?