

UNIVERSALITY OF MULTI-LAYER PERCEPTRON

INTEGRAL REPRESENTATION AND APPROXIMATION BOUND

Noboru Murata

June 20, 2023

<https://noboru-murata.github.io/>

Introduction

- mathematical model of neuron

- artificial neural network

Problem Formulation

- universality of three-layered perceptron

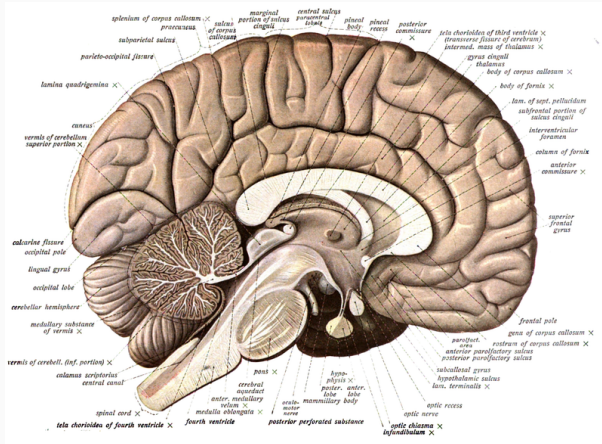
- approximation bound

- approximation error

Conclusion

INTRODUCTION

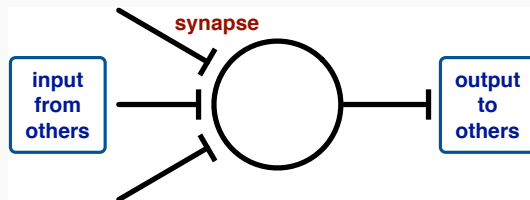
SPECIFICATION OF BRAIN



An anatomical illustration from Sobotta's Human Anatomy 1908

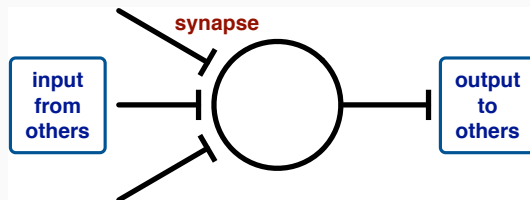
- weight: 1400g (2-3% of body)
- neurons:
 - cerebrum – 1.4×10^{10}
 - cerebellum – 1.0×10^{11}
- neuroglia:
ten times of neurons
- synapses:
 $10^3 - 10^5$ per neuron
- energy consumption:
 - blood – 15%
 - oxygen – 20%
 - dextrose – 25%

output



- output: pulses from 0Hz to 500Hz
- normalize
 - max frequency: 500Hz \mapsto 1
 - min frequency: 0Hz \mapsto 0

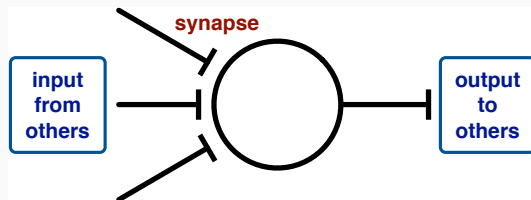
internal state



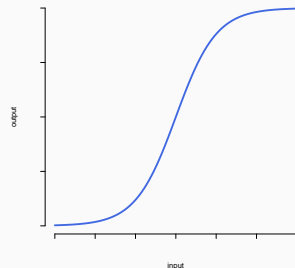
- input from other neuron: x_i
- strength of synapse: w_i
- internal state: **weighted sum of inputs**

$$u = \sum_i w_i x_i$$

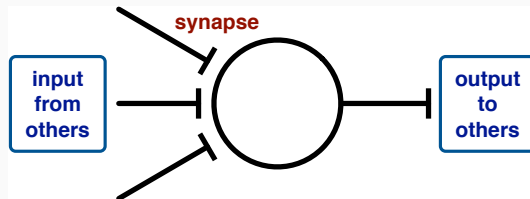
activation



- output a pulse when the internal state exceeds a certain constant:
thresholding
- range from 0 to 1:
non-linear transformation



activation input-output

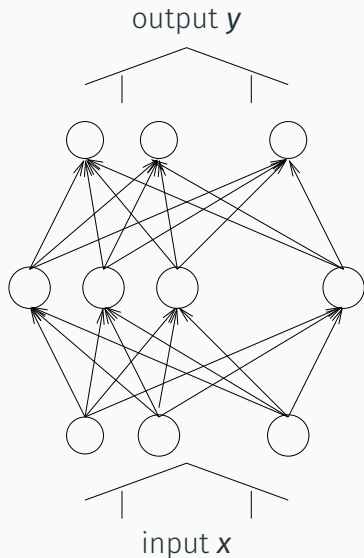


$$y = \psi \left(\sum_{i=1}^m w_i x_i - \theta \right) \quad (\text{model of a neuron})$$

y : output

θ : threshold

ψ : activation function



a simple calculation system consists of mathematical neurons

$$y_i = \sum_{j=1}^h c_{ij} \psi \left(\sum_{k=1}^m a_{jk} x_k - b_j \right),$$

$$(i = 1, \dots, l)$$

(m -dim input, 1-dim output)

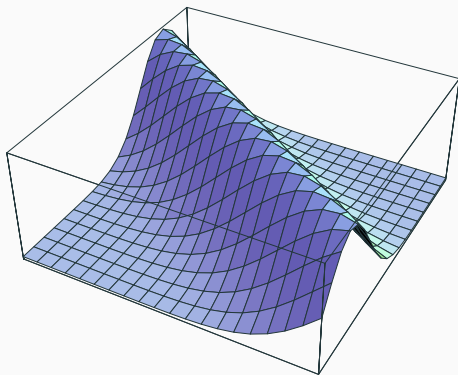
- easily implemented on computers because of homogeneously structured simple units
- simple and fast learning algorithms
(error-backpropagation: gradient method calculated via chain rule)
- size of units and structure of network can be roughly designed without detailed prior knowledges
- learning from examples sometimes gives a unexpected result, which may include important information of data inside networks

PROBLEM FORMULATION

Question

Find which class of functions can be well approximated by three layered perceptron with m -dim input and 1-dim output:

$$y = \sum_{j=1}^h c_j \psi \left(\sum_{k=1}^m a_{jk} x_k - b_j \right).$$



a ridge function on R^2

Definition (ridge function)

A function which is described with a vector $\mathbf{a} \in R^m$, a scalar $b \in R$, and a function $G : R \rightarrow R$ as

$$F(\mathbf{x}) = G(\mathbf{a} \cdot \mathbf{x} - b)$$

is called **ridge function**.

- suppose two functions $\phi_d, \phi_c \in L^1(R) \cap L^2(R)$ are bounded, and the following integral exists:

where $\hat{\cdot}$ denotes Fourier transform.

- $$T(\mathbf{a}, b) = \frac{1}{(2\pi)^m} \int_{R^m} \phi_d(\mathbf{a} \cdot \mathbf{x} - b) f(\mathbf{x}) d\mathbf{x}$$

kernel for composition

(combination of sigmoid functions)

$$\phi_c(z) = c\{\psi(z+h) - \psi(z-h)\}, \quad (h > 0, c: \text{constant})$$

$$\psi(z) = \frac{1}{1 + \exp(-z)}$$

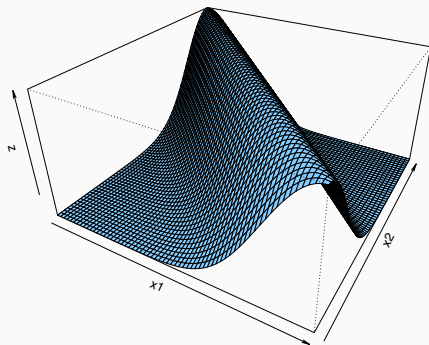
kernel for decomposition

(generalized differential operator)

$$\phi_d(z) = \begin{cases} c \frac{d^m}{dz^m} \rho(z) & m: \text{even} \\ c \frac{d^{m+1}}{dz^{m+1}} \rho(z) & m: \text{odd} \end{cases}$$

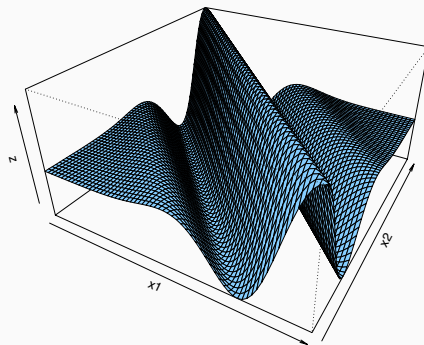
$$\rho(z) = \begin{cases} e^{-1/(1-|z|^2)} & |z| < 1 \\ 0 & |z| \geq 1 \end{cases}$$

$$z = \phi_c(x)$$



kernel for composition: ϕ_c

$$z = \phi_d(x)$$



kernel for decomposition: ϕ_d
(differential operator)

Theorem (NM 1996)

With transform T

$$T(\mathbf{a}, b) = \frac{1}{(2\pi)^m} \int_{R^m} \phi_d(\mathbf{a} \cdot \mathbf{x} - b) f(\mathbf{x}) d\mathbf{x},$$

function f is represented by

$$f(\mathbf{x}) = \lim_{\varepsilon \rightarrow 0} \int_{R^{m+1}} \phi_c(\mathbf{a} \cdot \mathbf{x} - b) T(\mathbf{a}, b) e^{-\varepsilon |\mathbf{a}|^2} d\mathbf{a} db.$$

If $f \in L^1(R^m) \cap L^p(R^m)$ ($1 \leq p < \infty$), the above equation converges in terms of L^p -norm. If $f \in L^1(R^m)$, bounded and uniformly continuous, the equation converges in terms of L^∞ -norm.

- define:

$$f_\varepsilon(\mathbf{x}) = \int_{\mathbb{R}^m} \int_{\mathbb{R}} \int_{\mathbb{R}^m} f(\mathbf{y}) \overline{\phi_d(\mathbf{a} \cdot \mathbf{y} - b)} \phi_c(\mathbf{a} \cdot \mathbf{x} - b) e^{-\varepsilon \|\mathbf{a}\|^2} d\mathbf{y} da db$$

- by Parseval's equality:

$$\int_{\mathbb{R}} \overline{\phi_d(\mathbf{a} \cdot \mathbf{y} - b)} \phi_c(\mathbf{a} \cdot \mathbf{x} - b) db = \int_{\mathbb{R}} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) e^{i\omega \mathbf{a} \cdot (\mathbf{x} - \mathbf{y})} d\omega$$

- thanks to the nature of Gaussian:

$$\begin{aligned}
 f_\varepsilon(\mathbf{x}) &= \int_{\mathbb{R}} \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) e^{i\omega \mathbf{a} \cdot (\mathbf{x} - \mathbf{y})} e^{-\varepsilon \|\mathbf{a}\|^2} f(\mathbf{y}) d\omega d\mathbf{y} d\mathbf{a} \\
 &= (2\pi)^m \int_{\mathbb{R}^m} G_{1/2\varepsilon}(\mathbf{a} - i\omega(\mathbf{x} - \mathbf{y})/2\varepsilon) d\mathbf{a} \\
 &\quad \int_{\mathbb{R}} \int_{\mathbb{R}^m} |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) G_{2\varepsilon/\omega^2}(\mathbf{x} - \mathbf{y}) f(\mathbf{y}) d\omega d\mathbf{y} \\
 &= (2\pi)^m \int_{\mathbb{R}} |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) G_{2\varepsilon/\omega^2} * f(\mathbf{x}) d\omega
 \end{aligned}$$

where

$$G_{\sigma^2}(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}^m} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right)$$

- by Hölder's inequality:

$$\begin{aligned}
 & \|f_\varepsilon - f\| \\
 &= \left\| (2\pi)^m \int_{\mathbb{R}} |\omega|^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) (G_{2\varepsilon/\omega^2} * f - f) d\omega \right\| \\
 &\leq (2\pi)^m \int_{\mathbb{R}} \left| \omega^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) \right| \left\| G_{2\varepsilon/\omega^2} * f - f \right\| d\omega \\
 &= (2\pi)^m \left[\int_{|\omega| \geq \gamma} + \int_{|\omega| < \gamma} \right] \\
 &\quad \left| \omega^{-m} \overline{\hat{\phi}_d(\omega)} \hat{\phi}_c(\omega) \right| \left\| G_{2\varepsilon/\omega^2} * f - f \right\| d\omega
 \end{aligned}$$

Question

Suppose a function f is represented by a transform T as

$$f(\mathbf{x}) = \int T(\mathbf{a}, b) \phi_c(\mathbf{x}; \mathbf{a}, b) d\mathbf{a} db.$$

Evaluate the accuracy of a finite sum of ϕ_c

$$f_n(\mathbf{x}) = \sum_i^n c_i \phi_c(\mathbf{x}; \mathbf{a}_i, b_i).$$

- a function f is represented by a transform T as

$$f(\mathbf{x}) = \int T(\mathbf{a}, b) \phi_c(\mathbf{x}; \mathbf{a}, b) d\mathbf{a} db.$$

- consider a finite sum of ϕ_c :

$$f_n(\mathbf{x}) = \sum_i^n c_i \phi_c(\mathbf{x}; \mathbf{a}_i, b_i).$$

- suppose inputs $\mathbf{x} \in R^m$ are generated subject to a probability density $\mu(\mathbf{x})$, evaluate the approximation by n units with the following norm:

$$\|f_n(\mathbf{x}) - f(\mathbf{x})\|_{L^2(R^m, \mu)}^2 = \int_{R^m} (f_n(\mathbf{x}) - f(\mathbf{x}))^2 \mu(\mathbf{x}) d\mathbf{x}$$

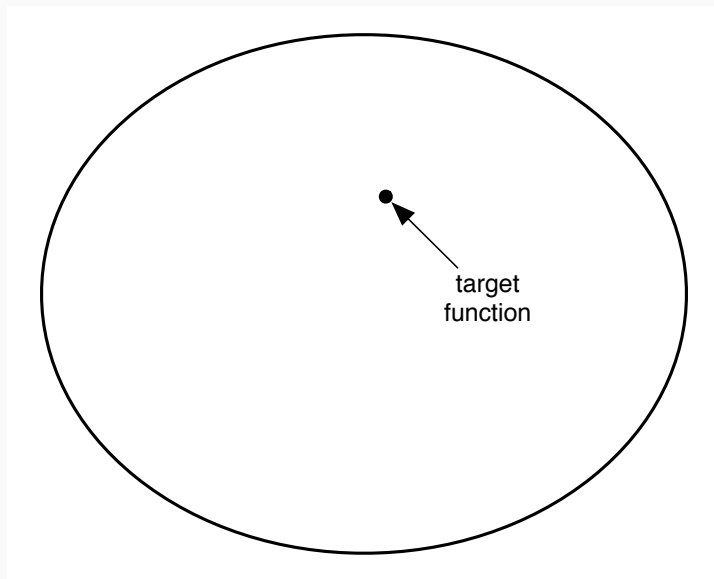
Theorem (NM 1996)

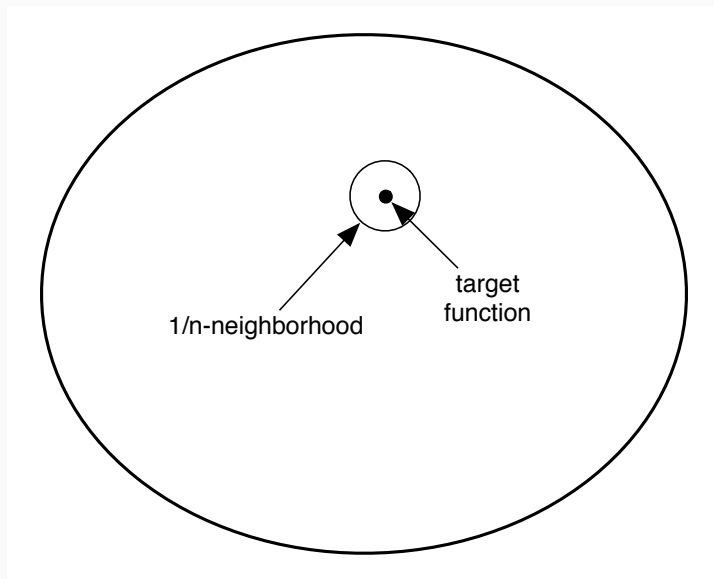
Suppose a function f is represented by a transform T as

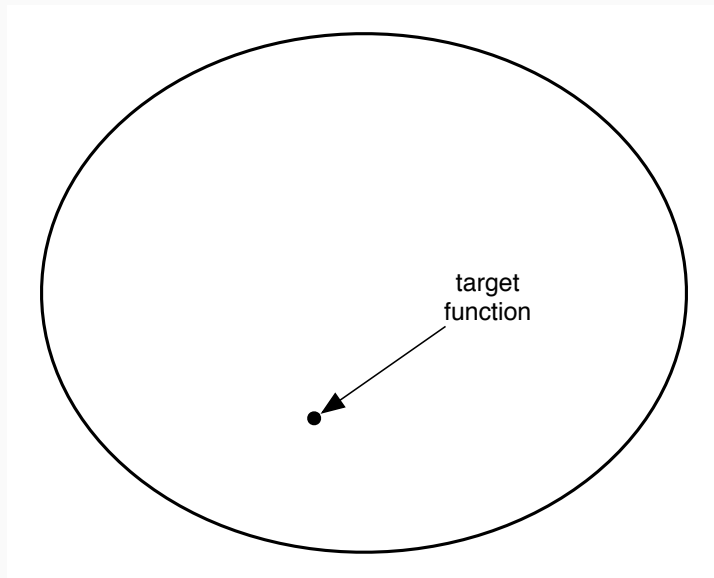
$$f(\mathbf{x}) = \int T(\mathbf{a}, b) \phi_c(\mathbf{x}; \mathbf{a}, b) d\mathbf{a} db.$$

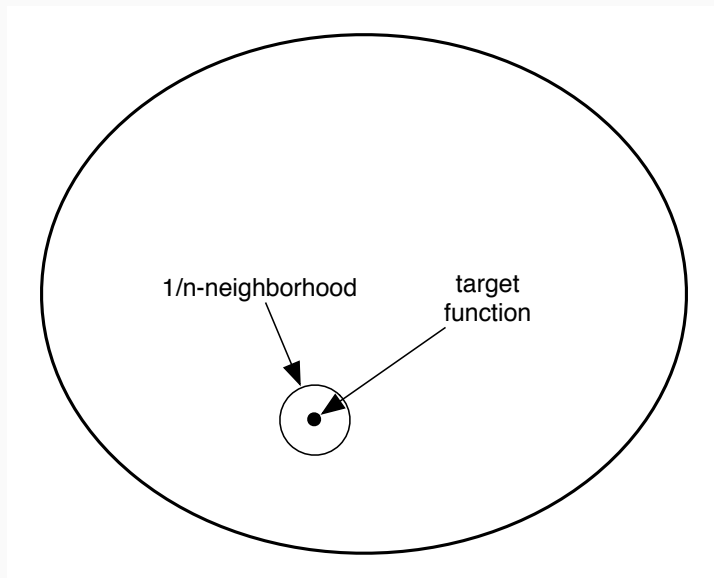
If the L_1 -norm (absolute integral) of T , $\|T\|_{L^1}$, is bounded, there exists an approximation f_n with a sum of n ϕ_c 's which satisfies

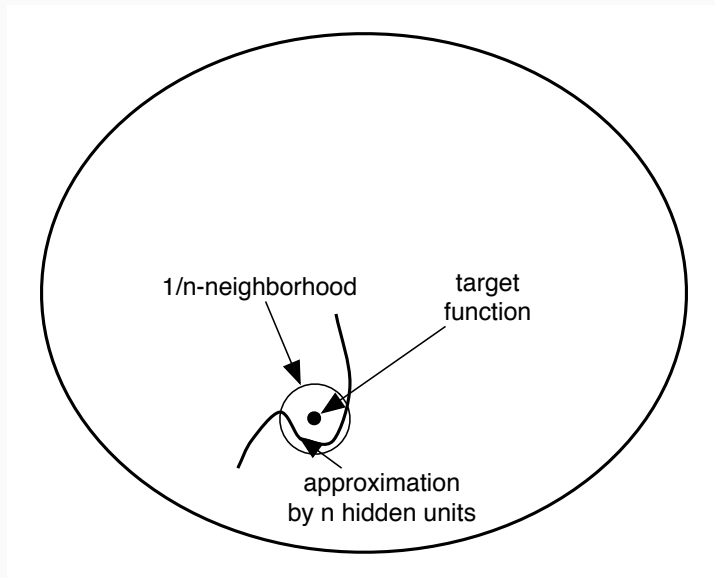
$$\|f_n(\mathbf{x}) - f(\mathbf{x})\|_{L^2(R^m, \mu)}^2 \leq \frac{1}{n} \|T\|_{L^1}^2.$$

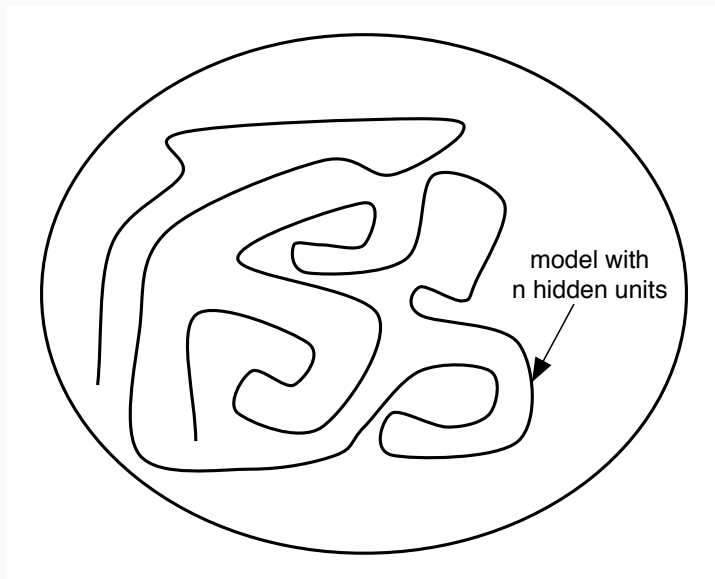












- for fixed \mathbf{x} , consider a random variable

$$X_i = c_i \phi_c(\mathbf{a}_i \cdot \mathbf{x} - b_i),$$

then

$$EX_i = f(\mathbf{x}), \quad V(X_i) \leq \|T\|_{L^1}^2 \cdot \left(\max_z \phi_c(z) \right)^2.$$

in the following discussion, assume $|\phi_c| < 1$.

- mean squared error of function f_n is evaluated as

$$\begin{aligned} E \int (f_n(\mathbf{x}) - f(\mathbf{x}))^2 \mu(\mathbf{x}) d\mathbf{x} &= \int V(f_n(\mathbf{x})) \mu(\mathbf{x}) d\mathbf{x} \\ &= \int V\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \mu(\mathbf{x}) d\mathbf{x} \leq \frac{1}{n} \|T\|_{L^1}^2. \end{aligned}$$

example of function spaces with $O(1/n)$ -rate convergence

function space	approximation
$\int \hat{f}(\omega) d\omega < \infty$	$\sum_{i=1}^n c_i \sin(\mathbf{a}_i \cdot \mathbf{x} - b_i)$ (Jones 1992)
$\int \omega \hat{f}(\omega) d\omega < \infty$	$\sum_{i=1}^n c_i \sigma(\mathbf{a}_i \cdot \mathbf{x} - b_i)$ (Barron 1993)
m -th Hölder continuous	$\sum_{i=1}^n c_i \sigma(\mathbf{a}_i \cdot \mathbf{x} - b_i)$ (NM 1996)
$H^{2p,1}(R^m)$, $2p > m$	$\sum_{i=1}^n c_i e^{- \mathbf{x} - \mathbf{a}_i ^2 / b_i^2}$ (Girosi 1993)

where σ is the sigmoid function, $H^{2p,1}(R^m)$ is the Sobolev space of $2p$ -th order differentiable.

- **aim:** minimize approximation errors of a contaminated function $y = f(\mathbf{x}) + \xi$
 - $f_{n,opt}$ – not obtainable

$$\text{minimize } \|y - f_n\|^2 = E_{\mathbf{x}, y}(y - f_n(\mathbf{x}))^2$$

- $f_{n,t}$ – obtainable

$$\text{minimize } \frac{1}{t} \sum_{j=1}^t (y_j - f_n(\mathbf{x}_j))^2$$

- error decomposition:

$$\|y - f_{n,t}\|^2 \Rightarrow \underbrace{\|y - f_{n,opt}\|^2}_{\text{structural error}} + \underbrace{\|f_{n,opt} - f_{n,t}\|^2}_{\text{learning error}}$$

- errors caused by training from examples:

$$E [\|y - f_{n,t}\|^2] = \|y - f_{n,opt}\|^2 + \frac{1}{2t} \text{tr}GH^{-1} + o\left(\frac{1}{t}\right)$$

$$V [\|y - f_{n,t}\|^2] = \frac{1}{2t^2} \text{tr}GH^{-1}GH^{-1} + o\left(\frac{1}{t^2}\right),$$

where ij -elements of G and H are given by using the partial derivative with respect to the i -th element, ∂_i , as

$$G_{ij} = E_{\mathbf{x},y}(\partial_i(y - f_n(\mathbf{x}))^2 \partial_j(y - f_n(\mathbf{x}))^2)$$

$$H_{ij} = E_{\mathbf{x},y}(\partial_i \partial_j (y - f_n(\mathbf{x}))^2).$$

CONCLUSION

we have investigated

- integral representation of three-layered perceptron
- approximation bounds of some function spaces

further works are done on

- specifying classes of activation functions
- investigating reproducing kernel Hilbert spaces

-  Murata, Noboru (Aug. 1996). "An Integral Representation of Functions Using Three-layered Networks and Their Approximation Bounds." In: *Neural Networks* 9 (6), pp. 947–956. DOI: [10.1016/0893-6080\(96\)00000-7](https://doi.org/10.1016/0893-6080(96)00000-7).