Azerbaijan Journal of Mathematics
V. 3, No 2, 2013, July
ISSN 2218-6816

Learning Theory and Approximation by Neural Networks

V. Maiorov

Abstract. This paper quantifies the approximation capability of Neural Networks and their application in machine learning theory. The problem of Learning Neural Networks from samples is considered. The sample size which is sufficient for obtaining the almost-optimal stochastic approximation of function classes is obtained. In the terms of the accuracy confidence function we show that the least square estimator is almost-optimal for the problem. Moreover, we consider the analogous problems related to learning by radial basis functions.
Learning theory is a growing field of research which attracts a large number of researchers from a variety of disciplines such as computer science, economics and neural networks. Mathematics is important for investigating learning problems since it provides the necessary level of rigorous analysis that leads to understanding the fundamental concepts and properties of learning. Specifically, the learning problem is reduced to finding a regression function (the average function of a given random processes) using the corresponding manifold under the condition that the function is not known but belongs to some given class of functions. The learning network problems have a long history (see the works of V. Vapnik, M.G.D. Powell, P.L. Bartlett, etc.)

Key Words and Phrases: Learning Theory, Neural Networks, Entropy, Stochastic approximation

2010 Mathematics Subject Classifications: 68Q32, 62L20, 68T05, 82C32

1. Introduction

Learning theory is a growing field of research which attracts a large number of researchers from disciplines such as physical or biological systems, engineering applications, financial studies. Mathematics is important for investigating learning problems since it provides the necessary level of rigorous analysis that leads to understanding the fundamental concepts and properties of learning. In many systems only finite number of data $(x_i, y_i)_{i=1}^m$ can be obtained. Leaning means synthesizing a function that represents the relation between the inputs and corresponding outputs. A learning system is normally developed for defining the function and yielding an estimator. The learning system comprises a hypothesis space, a family of parameterized functions that regulate the forms and properties of the estimator to be found, and a learning strategy or an learning algorithm that numerically yields the parameters of the estimator.

Specifically, the learning problem is reduced to finding a regression function (the average function of a given random processes) using the corresponding manifold under the condition that the function is not known but belongs to some given class of functions.

For results in other settings we recommend a book of V. Vapnik [34] , a survey by T. Evgeniou, M. Pontil and T. Poggio [9], and a survey on the classification problem and universal algorithms by G. Lugosi [16] and L. Devroy, L. Györfy, G. Lugosi [8] (see also the papers [4, 2, 25, 26]). Important results including a construction of universal algorithms of the learning theory are obtained in [6, 7, 14, 15, 32].

The learning neural network problem has a long history (see the works of Powell [30] and M. Anthony & P.L. Bartlett [1], dedicated to theoretical foundations of Neural Networks Learning. A. Pinkus [28] published the survey in the area of functional approximation by Neural Networks. S. Smale [31] set up the problem of the recovery of a target function given by a stochastic collection of samples using Neural Network manifolds. In this work we investigate that problem. In this connection, the results of the present paper are the continuation of [18, 19, 20, 17, 21, 22].

1.1. Learning Theory

1. One of the main problems of learning theory is to recover a function $y = f(x)$ having only some *a priori* information about it. For instance, having a finite collection of points $(x_1, y_1), ..., (x_m, y_m)$ as a sample of i.i.d. randomly drawn vectors (x, y) distributed according to some probability law which in general is not known.

Consider now the problem in detail. We adopt most of the notations from [4]. Let X and Y be some compact sets in spaces $\mathbb{R}^d$ and $\mathbb{R}$, respectively. Assume that on the product $X \times Y$ a Borel measure ρ is defined with $\rho(X \times Y) = 1$. Let $f : X \to Y$ be a function acting from X to Y. Denote by

$$\mathcal{E}(f) := E_\rho(f(x) - y) := \int_{X\times Y} (f(x) - y)^2 d\rho(x, y) \tag{1}$$

the expected value (or average) of the function $(f(x) - y)^2$ on $(X \times Y, \rho)$.

Let x be a fixed point in the set X. Denote by $\rho(y|x)$ the regular conditional measure on the set Y. Consider the average function on the set X

$$f_\rho(x) = \int_Y y\, d\rho(y|x),$$

called *the regression function* of the measure $\rho(y|x)$. Denote by

$$\sigma_\rho^2(x) = \int_Y (f_\rho(x) - y)^2 d\rho(y|x)$$

the dispersion function of the measure $\rho(y|x)$. The measure ρ induces the measure μ (or ρ_X) on the set X which is defined as $\mu(Q) = \rho(Q \times Y)$ for any measurable set $Q \subset X$. Averaging over X, we define

$$\sigma_\rho^2 := \int_X \sigma_\rho^2(x) d\mu(x).$$

2. Let $z_i = (x_i, y_i)$, $i = 1, ..., m$, be an arbitrary sample of m points in the set $Z = X \times Y$. Construct the empirical error of f corresponding to the point $z = (z_1, ..., z_m)$ of the set Z^m as

$$\mathcal{E}_z(f) := E_z[(f(x) - y)^2] = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

The quantity $\mathcal{E}_z(f)$ is the discrete analog of the average quantity (1).

3. Consider the Hilbert space $L_2(X, \mu)$ of functions defined on X. Let $W = W(X)$ be some class of functions in the space $L_2(X, \mu)$ with the norm

$$\|f\|_{X,\mu} = \left(\int_X |f(x)|^2 \, d\mu(x)\right)^{1/2}.$$

Assume that the measure ρ is such that the regression function f_ρ belongs to the class W. Let H be a compact manifold of functions in $L_2(X, \mu)$ which is called *the hypothesis space.* For a given vector $z = (z_1, .., .z_m)$ on Z^m we consider the function

$$f_z(x) = \mathrm{argmin}_{f \in H} \mathcal{E}_z(f) \tag{2}$$

in H, that is the function f_z, which achieves the minimum

$$\mathcal{E}_z(f_z) := \min_{f \in H} \mathcal{E}_z(f).$$

The function f_z is called *the least-squares estimator.* A simple calculation shows $\mathcal{E}(f_\rho) = \sigma_\rho^2$ and

$$\mathcal{E}(f_z) = \|f_z - f_\rho\|^2_{X,\mu} + \sigma^2_\rho.$$

We primarily measure the approximation error in the $L_2(X, \mu)$ space. If we have a particular least-squares approximant f_z to f_ρ, the quantity of the performance is measured by

$$\mathcal{E}(f_z) - \mathcal{E}(f_\rho) = \|f_z - f_\rho\|^2_{X,\mu}. \tag{3}$$

Define the measure $\rho^{\mathrm{m}} = \rho \times ... \times \rho$ (m times) on the set Z^m to be equal to the direct product of m copies of measure ρ. The error (3) clearly depends on z and therefore has a stochastic nature. As a result, it is impossible to say something about (3) in general for fixed z. Instead, we can look at its behavior in probability as measured by the expected error

$$E^{\mathrm{LS}}_{\rho^{\mathrm{m}}}(\|f_z - f_\rho\|^2_{X,\mu}) := \int_{Z^m} \|f_z - f_\rho\|^2_{X,\mu} \, d\rho^{\mathrm{m}}(z),$$

where the expectation is taken over all realizations z obtained for a fixed m.

From the law of large numbers it follows that by choosing suitable f_z, $E_{\rho^{\mathrm{m}}}(\|f_z - f_\rho\|^2_{X,\mu}) \to 0$ as $m \to 0$.

Consider the class $\mathbb{E}_m$ of all estimators, that is the class of all possible mappings $h_z : Z^m \to H$. Also we consider the expected error

$$E_{\rho^{\mathrm{m}}}(\|h_z - f_\rho\|^2_{X,\mu}) := \int_{Z^m} \|h_z - f_\rho\|^2_{X,\mu} \, d\rho^{\mathrm{m}}(z).$$

How fast $E_{\rho^{\mathrm{m}}}(\|h_z - f_\rho\|^2_{X,\mu})$ tends to zero depends at least on three things:

a) the nature of f_ρ,

b) the approximation properties of hypothesis space H,

c) how well we do in constructing the estimators h_z.

Let $\mathcal{M}$ be some class of Borel measures on X. Recall that we do not know ρ, so the best we can say about it is that $f_\rho \in W$. Given natural m, we define *the least square average accuracy m-deviation* as

$$e^{\mathrm{LS}}_m(W, H, \mathcal{M}) := \sup_\rho E_{\rho^{\mathrm{m}}}(\|f_z - f_\rho\|^2_{X,\mu}),$$

where f_z is *the least-squares estimator*, and ρ runs over all measures such that μ, which is the restriction of ρ on X, belongs to $\mathcal{M}$.

We enter into competition over all estimators h_z and define *the average accuracy m-deviation* as

$$e_m(W, H, \mathcal{M}) := \inf_{h_z} \sup_\rho E_{\rho^{\mathrm{m}}}(\|h_z - f_\rho\|^2_{X,\mu}),$$

where h_z runs over all possible estimators from the class $\mathbb{E}_m$, and ρ runs over all measures such that μ, which is the restriction of ρ on X, belongs to $\mathcal{M}$.

4. Let $m \in \mathbb{N}$ and $\varepsilon > 0$ be any numbers. We will study the following function (see [7],[33]) that is called *the accuracy confidence function*

$$\mathbf{AC}_m(W, H, \mathcal{M}, \varepsilon) = \inf_{h_z} \sup_\rho \rho^m(z : \ \|f_\rho - h_z\|_{X,\mu} \geq \varepsilon),$$

The mapping h_z which corresponds to the minimum is called *universal estimator* for the quantity $\mathbf{AC}_m(W, H, \mathcal{M}, \varepsilon)$.

Also we consider *the accuracy confidence function estimator for the least square method*

$$\mathbf{AC}^{\mathrm{LS}}_m(W, H, \mathcal{M}, \varepsilon) = \sup_\rho \rho^m(z : \ \|f_\rho - f_z\|_{X,\mu} \geq \varepsilon),$$

where we calculate the estimator f_z by the least square formula (2). Obviously, we have

$$\mathbf{AC}_m(W, H, \mathcal{M}, \varepsilon) \leq \mathbf{AC}^{\mathrm{LS}}_m(W, H, \mathcal{M}, \varepsilon). \tag{4}$$

The quantity $\mathbf{AC}_m(W, W, \mathcal{M}, \varepsilon)$, where the set W coincides with the hypothesis space H, was introduced by DeVore, Kerkyacharian, Picard and Temlyakov [6],[7]. In these works (see also [14, 15, 32]) optimal estimates for the accuracy confidence function estimator $\mathbf{AC}_m$ of a class W are obtained under condition that a behavior of the entropy (or Kolmogorov's) is given and has the order n^{-r}. Temlyakov [32] constructed the universal estimator f_z using the method of least square of the form

$$f_z = \mathrm{argmin}_{f \in \mathcal{N}_\varepsilon} \sum_{i=1}^m (f(x_i) - y_i)^2,$$

where f runs over all functions from a ε-net $\mathcal{N}_\varepsilon$ of the set W. A series of problems closely connected with given subjects are considered in [13, 9, 8, 16, 12]

In the present work we estimate (see Section 3) the quantities $\mathbf{AC}_m^{\mathrm{LS}}$ and $\mathbf{AC}_m$ for (defined below) the Sobolev class W_2^r of functions on the unit cube and and for the neural networks manifold $H = H_n^{NN}$. We obtain almost-optimal (with additional logarithmic factor) estimates of these quantities. Moreover we construct the universal estimator f_z using the standard method of least square, that is we construct the estimator by the formula

$$f_z = \mathrm{argmin}_{f \in H_n^{NN}} \sum_{i=1}^{m} (f(x_i) - y_i)^2,$$

where the minimum is calculated over all (not only ε-net) functions f from the hypothesis space H_n^{NN}. This circumstance permits us to apply the method of least square using parameters defining the space H_n^{NN}. The universality of the standard method of least squares, that is the lower bound for the quantity $\mathbf{AC}_m$, we obtain using the results of works [7], [33].

1.2. The Neural Network Manifold and Affine-Invariant Dictionary

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any sigmoidal function, that is σ is non-decreasing and $\lim_{t\to-\infty} \sigma(t) = 0$, $\lim_{t\to\infty} \sigma(t) = 1$. We consider *the neural network manifold* of functions

$$H_n^{NN}(\sigma) := \left\{ h(\bullet) = \sum_{k=1}^{n} c_k\, \sigma(a_k \cdot \bullet + b_k) :\ a_k \in \mathbb{R}^d,\ c_k,\, b_k \in \mathbb{R} \ \text{for all k} \right\}, \qquad (5)$$

where $a \cdot x$ is the inner product of a and x. The function σ is said to be *the generator function* of the neural network manifold. Consider in the manifold $H_n^{NN}(\varphi)$ the sub-manifold $H_n^{NN}(\varphi, M, \Omega)$ which consists of functions $h \in H_n^{NN}(\varphi)$ satisfying $|h(x)| \le M$ for all $x \in \Omega$.

Let $\mathcal{A}^d = \{(A, b)\}$ be the set of all affine mappings in the space $\mathbb{R}^d$ of the form $Ax + b$, $x \in \mathbb{R}^d$, where A and b run over the set of all real square matrices of order d and the set of all vectors in $\mathbb{R}^d$, respectively. We identify the set of the mappings $\mathcal{A}^d$ with the space $\mathbb{R}^{d^2+d}$. Let $\varphi(x)$ be any function from the space $L_2(\mathbb{R}, \mu)$. Consider the set of functions on $\mathbb{R}^d$

$$\mathcal{D}(\varphi) = \{\varphi(A \bullet + b) :\ (A, b) \in \mathcal{A}^d\},$$

which is called *affine-invariant dictionary.* Let n be any natural number. Using the dictionary, we generate the set of functions

$$H_n^{AI}(\varphi) = \left\{ h(\bullet) = \sum_{k=1}^{n} c_k \varphi(A_k \bullet + b_k) :\ c_k \in \mathbb{R},\ (A_k, b_k) \in \mathcal{A}^d \ \text{for all k} \right\}, \qquad (6)$$

consisting of all possible linear combinations of n functions from the dictionary $\mathcal{D}(\varphi)$. Note that $H_n^{NN}(\sigma)$ belongs to the affine-invariant manifold $H_n^{AI}(\varphi)$, where the function φ is defined as $\varphi(x_1, ..., x_d) = \sigma(x_1)$.

We will consider the sub-manifold $H_n^{AI}(\varphi, M, X)$ in $H_n^{AI}(\varphi)$ which consists of functions $h \in H_n^{AI}(\varphi)$ satisfying $|h(x)| \le M$ for all $x \in \Omega$. Let $K \ge 1$ be any number. Denote by

$H_n^{AI}(\varphi, M, X, K)$ the subset of all functions h in $H_n^{AI}(\varphi, M)$ such that all elements of all matrices $A_k = \{a_{ij}^k\}_{i,j=1}^d$ and all vectors $b_k = \{b_i^k\}_{i=1}^d$ and also all numbers c_k in (10) are bounded modulo by K, that is $|a_{ij}^k|, |b_i^k|, |c_k| \le K$ for all i, j and k.

The manifolds H_n^{NN} and H_n^{AI} play an important role in neural networks and learning theory. These manifolds are utilized for approximating functions which are learnt empirically from samples. A major problem here is the estimation of the asymptotic characteristics of the error of the empirical minimizer f_z. The error is often expressed in terms of the ε-entropy, Vapnik-Chervonenkis dimension and Pseudo-dimension of the manifolds. The concept of ε-entropy of a set is closely connected to the concepts VC-dimension (or Pseudo-dimension) of the set. Often the upper estimates for ε-entropy are obtained using an estimate of the VC-dimension of a given set (see Vapnik and Chervonenkis [35], Haussler [10], Mendelson and Vershinin [27]).

2. Structural and Approximation Properties of Neural Networks Manifold

Let $X = [-1, 1]^d$ and M be a positive number. Consider the set $Z = X \times [-M, M]$. Let ρ be a Borel probabilistic measure on the set Z, and $\mu = \rho(y|x)$ be a corresponding condition measure on the set X. Consider the space $L_2 := L_2(X, \mu)$ of all real square integrable functions on X with respect to the measure μ.

2.1. Entropy of the NN-manifolds

In this subsection we introduce the class Φ of generator functions and present known estimates for the ε-entropy and pseudodimension of the classes $H_n^{AI}(\varphi, M)$ and $H_n^{NN}(\varphi, M)$ with $\varphi \in \Phi$.

Let B be a Banach space and let H be a compact set in B. The quantity

$$\mathrm{Entr}_\varepsilon(H, B) = \log_2 N_\varepsilon(H, B),$$

where $N_\varepsilon(H, B)$ is the number of elements in the smallest ε-net of the set H, is called the *ε-entropy* of the set H in the space B. The quantity $N_\varepsilon(H, B)$ is called the *ε-covering number* of the set H.

Note that the estimation of the ε-entropy $\mathrm{Entr}_\varepsilon[H_n^{NN}(\sigma, M, X), L_2]$ of the class $H_n^{NN}(\sigma, M, X)$ is essentially dependent on the generator function σ. So from the paper [21] it follows that there exists a real-analytical sigmoidal function σ^* such that ε-entropy of the set $H_n^{NN}(\sigma^*, M, X)$ is equal to infinity for any $\varepsilon > 0$, $n \ge 3$, $M > 0$. Analogous statement holds for manifolds of a similar form $H_n^{AI}(\varphi, M, X)$. This fact poses a difficulty for statistical estimates which in general require a finite ε-entropy.

We will consider the generator functions φ having the form such that the ε-entropy of the manifolds $H_n^{AI}(\varphi, M, X)$ admits a finite value. Let $\mathcal{P}_s^d$ be the space of all real polynomial on d variables of degree at most s. We define the following four classes of functions:

1. The class $\Psi_s = \{\psi\}$ which consists of exponential functions of the form $\psi(x) = e^{p(x)}$, where $p \in \mathcal{P}_s^d$. For example, the Gaussian function $e^{-|x|^2}$, $|x|^2 = x_1^2 + ... + x_d^2$, belongs to the class Ψ_2.

2. The class $\Theta_s = \{\theta\}$ which consists of all rational functions of degree at most s, i.e. the functions of the form $\theta(x) = p(x)/q(x)$ where $p, q \in \mathcal{P}_s^d$, and $q(x) \neq 0$ for all $x \in \mathbb{R}^d$.

3. The class $\Lambda_s = \{\lambda\}$ which consists of all functions of the form

$$\lambda(x) = \frac{1}{1 + e^{p(x)}}, \qquad p \in \mathcal{P}_s^d.$$

For example, the standard function $\lambda(x) = 1/(1 + e^{-x_1})$ which belongs to the class Λ_1, is widely used in neural networks.

4. Let $G = G_{s_1,l}$ be any domain in the space $\mathbb{R}^d$ bounded by l polynomials surfaces of degree at most s_1, that is a domain of the form

$$G = \{x : \ p_i(x) \leq 0, \ i = 1, ..., l\},$$

where $p_1, ..., p_l$ are some polynomials from the space $\mathcal{P}_{s_1}^d$. We define the class $\Gamma_{s_1,s_2,l} = \{\gamma\}$ consisting of functions of the form $\gamma(x) = \chi_G(x)q(x)$, where χ_G is the characteristic function of the domain G, and q is a polynomial from the space $\mathcal{P}_s^d$. For example, the function $\gamma(x) = h(x_1)$, where $h(t) = \max\{t, 0\}$, $t \in \mathbb{R}$ is the Heaviside function, belongs to the class $\Gamma_{1,0,1}$.

Denote by Φ the class of functions which consists of the union of all functions from classes Ψ_s, Θ_s, Λ_s and $\Gamma_{s_1,s_2,l}$. Let $n \in \mathbb{N}$, $M > 0$ and $K \geq 1$ be fixed numbers. Consider the notation for manifolds

$$\mathcal{H}_n^{AI}(\varphi) = \begin{cases} H_n^{AI}(\varphi, M, 2X), & \text{if} \quad \varphi \in \Psi_s \\ H_n^{AI}(\varphi, M, X), & \text{if} \quad \varphi \in \Theta_s \\ H_n^{AI}(\varphi, M, X, K), & \text{if} \quad \varphi \in \Lambda_s \\ H_n^{AI}(\varphi, M, X), & \text{if} \quad \varphi \in \Gamma_{s_1,s_2,l}. \end{cases}$$

Henceforth we denote by c, c' and C_i, c_i, c_i', $i = 0, 1, ...$, the positive constants depending only on r, d, s, s_1, s_2, l, K and M. For two positive sequences a_n and b_n, $n = 0, 1, ...$ we write $a_n \asymp b_n$ if there exist positive constants c_1 and c_2 such that $c_1 \leq a_n/b_n \leq c_2$ for all $n = 0, 1, ...$ Also we denote $\log a = \log_2 a$.

Lemma 2.1. [17] *Let φ be any function from the class Φ. Then for any natural n and any positive number ε the following inequality holds*

$$\mathrm{Entr}_\varepsilon[\mathcal{H}_n^{AI}(\varphi), L_2(\mu)]\} \leq n\, T(\varepsilon, \varphi, d, M, K),$$

where

$$T(\varepsilon, \varphi, d, M, K) = \begin{cases} c_1 d^s \log^2 \frac{M}{\varepsilon}, & \text{if} \quad \varphi \in \Psi_s \\ c_2 d^2 \log n \log \frac{M}{\varepsilon}, & \text{if} \quad \varphi \in \Theta_s \\ c_3 d^2 \log n \log \frac{KM}{\varepsilon}, & \text{if} \quad \varphi \in \Lambda_s \\ c_4 d^2 \log n \log \frac{M}{\varepsilon}, & \text{if} \quad \varphi \in \Gamma_{s_1,s_2,l}. \end{cases}$$

Lemma 2.1 implies the following statement:

Consequence 2.2. *Let σ be the sigmoidal function such that the function $\varphi(x_1, ..., x_d) = \sigma(x_1)$ belongs to the class Φ. Denote $\mathcal{H}_n^{NN}(\sigma) = \mathcal{H}_n^{AI}(\varphi)$. Then the following inequality holds:*

$$\mathrm{Entr}_\varepsilon[\mathcal{H}_n^{NN}(\sigma), L_2(\mu)] \le n\, T(\varepsilon, \sigma, d, M, K).$$

2.2. Approximation by Neural Networks

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a sigmoidal function. Consider the neural network manifold $H_n^{NN}(\sigma)$ (see (5). In this section we state a theorem for the approximation of Sobolev class using the manifold $H_n^{NN}(\sigma)$.

Consider the space $L_2(X, \mu)$ of functions defined on $\mathbb{R}^d$ with support on the set X and the norm

$$\|f\|_{L_2(\mathbb{R}^d,\mu)} = \left(\int_X |f(x)|^2 \, d\mu(x)\right)^{1/2}.$$

Denote by BL_2 the unite ball in the space $L_2(\mathbb{R}^d, \mu)$. For any function $f \in L_2$ we denote by $\mathcal{F}(f)$ or $\hat{f}$ the Fourier transform of f

$$\hat{f}(u) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x)\, e^{iu\cdot x}\, dx,$$

where $u \in \mathbb{R}^d$. The inverse Fourier transform will be denoted by $\mathcal{F}^{-1}$. In the space L_2 define the derivative of order α as

$$\mathcal{D}^\alpha f := \mathcal{F}^{-1}\{|u|^\alpha \mathcal{F}(u)\},$$

where $|u| = \sqrt{u_1^2 + \cdots + u_d^2}$, and Fourier transform and derivatives are understood in the distribution sense. Let r be any positive number. In the space $L_2 = (X, \mu)$ consider the Sobolev class of functions

$$W_2^r := W_2^r(X, \mu) := \{f : \max_{0\le\alpha\le r} \|\mathcal{D}^\alpha f\|_{L_2(X,\mu)} \le 1\}.$$

Introduce the class Ψ^r of functions ψ satisfying following conditions. Let ψ be some function in the space $L_1(\mathbb{R})$. Using the function ψ, we construct an adjoint function φ on $\mathbb{R}$ satisfying the equality

$$\int_0^\infty a^{-1}\hat{\psi}(aw)\bar{\hat{\varphi}}(aw)\, da = 1 \tag{7}$$

for any w, where $\hat{\psi}$ and $\hat{\varphi}$ are the Fourier transforms of ψ and φ, respectively. The function class Ψ^r, where $r > 0$, consists of all functions $\psi \in L_2(\mathbb{R}) \bigcap L_1(\mathbb{R})$ for which there exists a function ψ satisfying (7) and such that for all $\rho \in [0, r]$, $\mathcal{D}^\rho\psi \in L_2(\mathbb{R})$ and $\mathcal{D}^{-\rho}\varphi \in L_1(\mathbb{R})$. We set

$$B_\varphi = \max_{0\le\rho\le r}\{\|\mathcal{D}^\rho\psi|_{L_2(\mathbb{R})}, \, \|\mathcal{D}^{-\rho}\varphi\|_{L_1(\mathbb{R})}\}. \tag{8}$$

Examples: One can easily verify that the functions

$$\psi(t)=\sqrt{2}e^{-t^2/2},\quad \frac{t+1}{3}\chi_{[-1,0]}(t)+\frac{1-t}{3}\chi_{[0,1]}(t),$$

where χ_Δ is the characteristic function of the segment Δ, belong to the class Ψ^r. The corresponding adjoint functions φ are respectively

$$\varphi(t)=\sqrt{2}(1-t^2)e^{-t^2/2},\quad -\chi_{[-1,0]}(t)+\chi_{[0,1]}(t).$$

We observe that in many neural networks applications it is usual to use sigmoidal functions which approach a constant non-zero value at infinity. However, by taking suitable linear combinations of such functions, one can always obtain functions vanishing at infinity which belongs to $L_1(\mathbb{R})$. For example, for the sigmoidal function $\sigma(t)$ we require $\sigma(t+1)-\sigma(t)=\varphi(t)$, and $\lim_{t\to-\infty}\varphi(t)=\lim_{t\to\infty}\varphi(t)=0$.

Theorem 2.3. ([18],[19]) *Let the measure $\mu\in\mathcal{M}^*$ and σ be any sigmoidal function such that the function $\psi(t)=\sigma(t+1)-\sigma(t)$ belongs to the class Ψ^r. Then for any function $f\in W_2^r(X,\mu)$ there exists a function $h\in H_n^{NN}(\sigma)$ such that*

$$1.\qquad \|f-h\|_{L_2(X,\mu)}\le\frac{c_1B_\varphi\ln n}{n^{r/d}},$$

$$2.\qquad |h(x)|\le c_2B_\psi n^{(1/2-r/d)_+}\ for\ all\ x\in X.$$

where $(t)_+=\max\{t,0\}$, and c_1, c_2 are some constants depending only on r and d.

In [18] the following result was proved: Let σ be any sigmoidal function such that the function $\psi(t)=\sigma(t+1)-\sigma(t)$ belongs to the class Ψ^r. Then for any function $f\in W_2^r$ there exists a function $h\in H_n^{NN}(\sigma)$ such that

$$\|f-h\|_{L_2(X,\mu)}\le\frac{c_1B_\varphi\ln n}{n^{r/d-\delta}},\tag{9}$$

where δ is any number of the form $\delta=\delta'+k\varepsilon/d$, $\delta'\in(0,1)$, $\varepsilon\ge n^{-\delta'}$ and k is the least natural number satisfying $r\le\frac{kd}{2}+k\varepsilon$. Note that the inequality (9) is proved in [18] for the case of Lebesgue measure μ. In the common case, taking into consideration the property of the measure $0<C_1\le d\mu/dx\le C_2$, $x\in X$, the inequality (9) also holds. From the inequality (9) the statement 1 follows. Indeed, we set $\varepsilon=\frac{1}{\ln n}$, $\delta=\frac{(\ln 1/\varepsilon)}{\ln n}$. Then from (9) we obtain

$$\|f-h\|_{L_2(X,\mu)}\le\frac{c_1B_\psi n^{\delta+k\varepsilon/d}}{n^{r/d}}=\frac{c_1B_\psi\ln n}{n^{r/d}}.$$

The second inequality in Theorem 2.3 is proved in ([19], Th. IV.1).

3. Estimates of the accuracy confidence function estimator and the average accuracy m-deviation

We consider the measure class $\mathcal{M}^*$ which consists of all measures μ defined on X such that the Radon-Nikodym derivative satisfies the inequalities $C_1 \le d\mu/dx \le C_2$ for all $x \in \Omega$, and C_1, C_2 are absolute positive constants. Also we require that the generator function belongs to the class $\Phi \bigcap \Psi^r$ (for the definitions of the classes Φ and Ψ^r see Subsection 2.1 and 2.2, respectively). In our work we obtain asymptotic two-sided estimates for the accuracy confidence functions

$$\mathbf{AC}^{\mathrm{LS}}_{m,n} := \mathbf{AC}^{\mathrm{LS}}_m(W^r_2, H^{NN}_n(\varphi, M), \mathcal{M}^*, \varepsilon),$$

and

$$\mathbf{AC}_{m,n} := \mathbf{AC}_m(W^r_2, H^{NN}_n(\varphi, M), \mathcal{M}^*, \varepsilon).$$

Theorem 3.1. ([23]) Let $r > d/2$ and φ be any function from the class $\Phi \bigcap \Psi^r$. Let $\in \mathbb{N}$, $\varepsilon > 0$ be any numbers, and $n(\varepsilon) = [c_0 \varepsilon^{-r/d} \log^{d/r} 1/\varepsilon]$. Then there exist the positive constants c_i, $i = 0, ..., 4$ and also $\varepsilon_0 > 0$ and ε^-_m, ε^+_m satisfying

$$\frac{c_1}{m^{r/(2r+d)}} \le \varepsilon^-_m \le \varepsilon^+_m \le \frac{c_2 \log m}{m^{r/(2r+d)}},$$

such that

$$\mathbf{AC}^{\mathrm{LS}}_{m,n(\varepsilon)} \ge \mathbf{AC}_{m,n(\varepsilon)} \ge \varepsilon_0, \quad \text{for} \quad \varepsilon < \varepsilon^-_m,$$

and

$$e^{-c_3 m \varepsilon^2} \le \mathbf{AC}_{m,n(\varepsilon)} \le \mathbf{AC}^{\mathrm{LS}}_{m,n(\varepsilon)} \le e^{-c_4 m \varepsilon^2}, \quad \text{for} \quad \varepsilon \ge \varepsilon^+_m.$$

From Theorem 4.1 it follows that the least square method is almost-optimal (with additional logarithmic factor) for the accuracy confidence function $\mathbf{AC}_m(W^r_2, H^{NN}_{n(\varepsilon)}(\varphi, M), \mathcal{M}^*, \varepsilon)$, that is, it realizes the universal algorithm for the function.

Theorem 3.2. *Let $r > d/2$ and φ be any function from the class $\Phi \bigcap \Psi^r$. Then there exist the positive constants c_5, c_6 such that*

$$\begin{aligned} c_5 m^{-\frac{2r}{2r+d}} &\le e_m(W^r_2, H^{NN}_m(\varphi, M), \mathcal{M}^*) \\ &\le e^{\mathrm{LS}}_m(W^r_2, H^{NN}_m(\varphi, M), \mathcal{M}^*) \le c_6 m^{-\frac{2r}{2r+d}} \ln^2 m. \end{aligned}$$

Theorem 3.2 shows that the least square method is almost-optimal (with additional logarithmic factor) for the average accuracy m-deviation, that is, it realizes the universal algorithm.

4. Learning by Radial Basis Function Networks

Consider the Euclidian norm $\|x\| = (\sum_{i=1}^{d} x_i^2)^{1/2}$ in the space $\mathbb{R}^d$. Let $\varphi(t)$ be a function defined on $\mathbb{R}_+$. Consider the radial function $\varphi(\|x\|)$ on $\mathbb{R}^d$.

In this section we consider the problem of Leaning by Radial Basis Function Networks, that is we estimate the accuracy confidence function estimator and the average accuracy m-deviation, while when as the hypothesis space we consider the space

$$H_n^{RBF}(\varphi) = \left\{ h(\| \bullet \|) = \sum_{k=1}^{n} c_k \varphi(\| \bullet + a_k \|) : \ c_k \in \mathbb{R}, \ a_k \in \mathbb{R}^d \right\}, \qquad (10)$$

consisting of all possible linear combinations of n radial basis functions .

Note that $H_n^{RBF}(\varphi)$ coincides with the affine-invariant manifold $H_n^{AI}(\varphi)$, where the function φ is defined as $\varphi(x) = \|x\|$ and the dictionary is the set $\mathcal{A}^d = (I, \mathbb{R}^d)$, where I is the unit matrix. From here and from [24], the results below follow.

Theorem 4.1. *Let $r > d/2$, $m \in \mathbb{N}$, $\varepsilon > 0$ be any numbers, and $n(\varepsilon) = [c_0 \varepsilon^{-r/d} \log^{d/r} 1/\varepsilon]$. Let φ be any univariate polynomial of degree $n(\varepsilon)$. We set*

$$\mathbf{AC}_{m,n(\varepsilon)}^{\mathrm{LS}} := \mathbf{AC}_m^{\mathrm{LS}}(W_2^r, H_{n(\varepsilon)}^{RBF}(\varphi, M), \mathcal{M}^*, \varepsilon),$$

and

$$\mathbf{AC}_{m,n(\varepsilon)} := \mathbf{AC}_m(W_2^r, H_{n(\varepsilon)}^{RBF}(\varphi, M), \mathcal{M}^*, \varepsilon).$$

Then there exist the positive constants c_i, $i = 0, ..., 4$ and also $\varepsilon_0 > 0$ and ε_m^-, ε_m^+ satisfying

$$\frac{c_1}{m^{r/(2r+d)}} \le \varepsilon_m^- \le \varepsilon_m^+ \le \frac{c_2 \log m}{m^{r/(2r+d)}},$$

such that

$$\mathbf{AC}_{m,n(\varepsilon)} \quad \ge \quad \mathbf{AC}_{m,n(\varepsilon)}^{\mathrm{LS}} \ge \varepsilon_0, \qquad \text{if } \varepsilon < \varepsilon_m^-,$$

and

$$e^{-c_3 m \varepsilon^2} \le \mathbf{AC}_{m,n(\varepsilon)} \le \mathbf{AC}_{m,n(\varepsilon)}^{\mathrm{LS}} \le e^{-c_4 m \varepsilon^2}, \qquad \text{if } \varepsilon \ge \varepsilon_m^+.$$

From Theorem 4.1 it follows that the least square method is almost-optimal (with additional logarithmic factor) for the accuracy confidence function $\mathbf{AC}_m(W_2^r, H_{n(\varepsilon)}^{NN}(\varphi, M), \mathcal{M}^*, \varepsilon)$, that is, it realizes the universal algorithm for the function.

Theorem 4.2. *Let $r > d/2$, $n \in \mathbb{N}$ and φ be any univariate polynomial of degree n. Then there exist the positive constants c_5, c_6 such that*

$$\begin{aligned} c_5 n^{-\frac{2r}{2r+d}} \quad &\le \quad e_n(W_2^r, H_n^{RBF}(\varphi, M), \mathcal{M}^*) \\ &\le \quad e_n^{\mathrm{LS}}(W_2^r, H_n^{RBF}(\varphi, M), \mathcal{M}^*) \le c_6 n^{-\frac{2r}{2r+d}} \ln^2 n. \end{aligned}$$

Theorem 3.2 shows that the least square method is almost-optimal (with additional logarithmic factor) for the average accuracy m-deviation, that is, it realizes the universal algorithm.

References

[1] M. Anthony and P.L. Bartlett, *Neural Network Learning, Theoretical Foundations*, Cambridge University Press, 1999.

[2] P.L. Bartlett, O. Bousquet and S. Mendelson, *Local Rademacher complexity*, Ann. Statist., to appear, 2005.

[3] M.S. Birman and M.Z. Solomyak, *Piecewise polynomial approximations of functions of the classes* W_p^α, Mat. Sb. **73** (1967), 331-355; English translation in *Math. USSR Sb* (1967), 295-317. MR **36**:576.

[4] F. Cucker and S. Smale, *On the mathematical foundations on learning*, Bull. of Amer. Math. Soc., **39** (2001), 1-49.

[5] I. Daubechies, *Ten lectures on wavelets*, SIAM Press, 1992.

[6] R. DeVore, G. Kerkyacharian, D.Picard and V. Temlyakov, *On Mathematical Methods of Learning*, IMI Preprints, **10** (2004), 1-24.

[7] R. DeVore, G. Kerkyacharian, D.Picard and V. Temlyakov, *Mathematical methods for supervised learning*, IMI Preprints, **22** (2004), 1-51.

[8] L. Devroy, L. Györfy, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, (1996).

[9] T. Evgeniou, M. Pontil and T. Poggio, *Regularization Network and Support Vector Mashines*, Advances in Comput. Math. **13** (2000), 1-50.

[10] D. Haussler, *Sphere Packing Numbers for Subset of the Boolean n-Cube with Bounded Vapnik-Chervonenkis Dimension*, J. of Combinatorical Theory, **69** (1995), 217-232.

[11] S. Helgason, *The Radon Transform*, Birkhäuser, Boston, 1980.

[12] L. Györfy, *Principles of nonparametric learning*, International centre for mechanical sciences, Courses and Lectures, No.434.

[13] L. Györfy, M. Kohler, A. Krzyzak and H. Walk, *A distribution-free theory of nonparametricregression*, Springer, Berlin, 2002.

[14] S. Konyagin and V. Temlyakov, *Some error estimates in Learning Theory*, IMI Preprints, **05** (2004), 1-18.

[15] S. Konyagin and V. Temlyakov, *The entropy in Learning Theory, Error estimates*, IMI Preprints, **09** (2004), 1-25.

[16] G. Lugosi, Pattern classification and learning theory, In Principles of Nonparametric Learning, Springer, Vena (2002), 5-62.

[17] V. Maiorov, Pseudo-dimension and entropy of manifolds formed by affine-invariant dictionary, Submitted to Advanced in Computational Mathematics.

[18] V. Maiorov and R. Meir, *On the near optimality of the stochastic approximation of smooth functions by neural networks*, Advanced in Computational Mathematics **13** (2000), 79-103.

[19] V. Maiorov and R. Meir, *On the optimality of networks approximation using incremental algorithms*, IEEE Transactions on neural networks, **11** (2000), 323-337.

[20] V. Maiorov and R. Meir, *Lower Bounds for Multivariate Approximation by Affine-Invariant Dictionaries*, IEEE Transactions on Information Theory, **47** (2001), 1569-1575.

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

[22] V. Maiorov and J. Ratsaby, *On the Value of Partial Information for Learning from Examples*, J.Complexity, **13** (1997), 509-544.

[23] V. Maiorov, Approximation by neural networks and learning theory. *J. of Complexity* 22 (2006), 102-117.

[24] V. Maiorov, Representations of polynomials by linear combination of radial basis functions,*Constr. Approx.*, **37** (2013), 283-293.

[25] S. Mendelson, *Learning Relatively Small Classes*, Proceedings of the 14-th annual conference on Computational Learning Theory, (2001), 273-288.

[26] S. Mendelson, *Geometric Methods in the Analysis of Glivenko-Cantelli Classes*, Proceedings of the 14th annual conference on Computational Learning Theory, (2001), 256-272.

[27] S. Mendelson and R. Vershinin, *Entropy and the Combinatorial Dimension*, Inventiones Mathematicae, **152** (2003),37-55.

[28] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numerica, Cambridge University Press, (1999), 1-52.

[29] T. Poggio and S. Smale, *The Mathematics of Learning: Dealing with Data*, Manuscript (2003), 1-16.

[30] M.J.D. Powell, *The of radial basis approximation in 1990*, In: Advances in Numerical Analysis, 2(W.A. Light, ed.). Oxford: Oxford University Press, pp. 105-210.

[31] S. Smale, *Mathematical Problems for the next century, Mathematics: Frontiers and Perspectives* (V. Arnold, M. Atiyah, P. Lax, B. Vfzur, eds.), AMS, 2000, 271-294. CMP 2000:13.

[32] V. N. Temlyakov, *Optimal Estimators in Learning Theory*, IMI Preprints, **23** (2004), 1-29.

[33] V. N. Temlyakov, *Approximation in Learning Theory*, IMI Preprints, **5** (2005), 1-43.

[34] V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, 1998.

[35] V. Vapnik, A. Chervonenkis, *Necessary and sufficient conditions for the uniform convergence of empirical means to their expectations*, Theory Probab. Appl. **3** (1981), 532-553.

V. Maiorov
Technion I.I.T., 32000 Haifa, Israel
E-mail: maiorov@tx.technion.ac.il