Industrial Mathematics Institute

2004:10

On mathematical methods of learning

R. DeVore, G. Kerkyacharian, D. Picard and V.Temlyakov

IMI
Preprint Series

Department of Mathematics
University of South Carolina

ON MATHEMATICAL METHODS OF LEARNING

R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov

1. Introduction

We discuss in this paper some mathematical aspects of supervised learning theory. Supervised learning, or learning-from-examples, refers to a process that builds on the base of available data of inputs x_i and outputs y_i, $i = 1, \dots, m$, a function that best represents the relation between the inputs $x \in X$ and the corresponding outputs $y \in Y$. The central question is how well this function estimates the outputs for general inputs. The standard mathematical framework for the setting of the above learning problem is the following ([CS], [PS]).

Let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ be Borel sets and let ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \to Y$ define *the error*

$$\mathcal{E}(f) := \mathcal{E}_\rho(f) := \int_Z (f(x) - y)^2 d\rho.$$

Consider $\rho(y|x)$ - conditional (with respect to x) probability measure on Y and ρ_X - the marginal probability measure on X (for $S \subset X$, $\rho_X(S) = \rho(S \times Y)$). Define

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

The function f_ρ is known in statistics as the *regression function* of ρ. It is clear that f_ρ minimizes the error $\mathcal{E}(f)$ over all $f \in L_2(\rho_X)$: $\mathcal{E}(f_\rho) \le \mathcal{E}(f)$, $f \in L_2(\rho_X)$. Thus, in the sense of error $\mathcal{E}(\cdot)$ the regression function f_ρ is the best to describe the relation between inputs $x \in X$ and outputs $y \in Y$. Now, our goal is to find an estimator f_z, on the base of given data $z = ((x_1, y_1), \dots, (x_m, y_m))$ that approximates f_ρ well with high probability. We assume that (x_i, y_i), $i = 1, \dots, m$ are independent and distributed according to ρ. There are several important ingredients in mathematical formulation of this problem. In our formulation we follow the way that has become standard in approximation theory and based on the concept of *optimal method*. A classical example of such a setting is the concept of the Kolmogorov width. Kolmogorov's n-width for centrally symmetric compact set F in Banach space B is defined as follows

$$d_n(F, B) := \inf_L \sup_{f \in F} \inf_{g \in L} \|f - g\|_B$$

where $\inf_L$ is taken over all n-dimensional subspaces of B. In other words the Kolmogorov n-width gives the best possible error in approximating a compact set F by n-dimensional linear subspaces. So, first of all we need to make an assumption on the unknown function f_ρ. Following the approximation theory approach we make this assumption in the form $f_\rho \in W$, where W is a given class of functions. For instance, we may assume that f_ρ has some smoothness. The next step is to find an algorithm for constructing an estimator f_z that is optimal for the class W. By optimal we mean the one that provides the minimal error $\|f - f_z\|$ for all $f \in W$ with high probability. A problem of optimization is naturally broken into two parts: upper estimates and lower estimates. We discuss only upper estimates in this paper. In order to prove upper estimates we need to decide what should be the form of an estimator f_z. In other words we need to specify the *hypothesis space* $\mathcal{H}$ (see [CS], [PS]) where an estimator f_z comes from. We may also call $\mathcal{H}$ an *approximation space.*

The next question is how to build $f_z \in \mathcal{H}$. In Section 2 we will discuss the method that takes

$$f_{z,\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_z(f),$$

where

$$\mathcal{E}_z(f) := \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$$

is the *empirical error* of f. This $f_{z,\mathcal{H}}$ is called the *empirical optimum.* Section 2 contains a discussion of known results from [CS] and some new results. Proofs of new results in Section 2 are based on the technique developed in [CS].

In Section 3 we assume that ρ is an absolutely continuous measure with density $\mu(x)$: $d\rho = \mu dx$. We study estimation of a new function $f_\mu := f_\rho\mu$ instead of regression function f_ρ. As a part of motivation of this new setting we discuss one practical example from finance. Let $x = (x^1, \dots, x^d) \in \mathbb{R}^d$ be information that is used by a bank to decide to give or not to give a mortgage to a client. For instance, x^1 - income, x^2 - home value, x^3 - mortgage value, x^4 - interest rate. Let y be the total profit (loss if negative) that the bank gains from this mortgage. Then $f_\rho(x)$ stands for the expected (average) profit of the bank from clients with the same information x. A clear goal of the bank is to find $S \subset \mathbb{R}^d$ that maximizes

$$\int_S f_\rho(x)d\rho_X.$$

Obviously, such S is given by

$$S_o = \{x : f_\rho(x) \geq 0\}.$$

The mathematical question in this regard is how to utilize the available data
$z = ((x_1, y_1), \dots, (x_m, y_m))$ to find an empirical S_z that gives a good approximation to $\int_{S_o} f_\rho(x)d\rho_X$.

We suggest the following way to solve this problem. Assume that $d\rho_X = \mu dx$ and denote $f_\mu := f_\rho\mu$. Then

$$\int_S f_\rho(x)d\rho_X = \int_S f_\mu(x)dx.$$

So, we look for an estimator for f_μ instead of f_ρ. In the above example it is sufficient to estimate $\|f_\mu - f_z\|_{L_1}$. Suppose we have found f_z such that

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu - f_z\|_{L_1} \le \epsilon\} \ge 1-\delta.$$

Define

$$S_z := \{x : f_z(x) \ge 0\}.$$

Then with the above estimate on the probability we have

$$\int_{S_z} f_\rho(x)d\rho_X = \int_{S_z} f_\mu(x)dx \ge \int_{S_z} f_z(x)dx - \epsilon$$

$$\ge \int_{S_o} f_z(x)dx - \epsilon \ge \int_{S_o} f_\rho(x)d\rho_X - 2\epsilon.$$

Therefore, the empirical set S_z provides an optimal profit within an error 2ϵ with probability $\ge 1-\delta$.

In the above example it was convenient to measure the error in the L_1 norm. However, it is usually simpler to estimate the L_2 error of approximation. We note that

$$\|f\|_{L_1} \le (\mathrm{mes}\, X)^{1/2}\|f\|_{L_2}, \quad \text{and} \quad \|f\|_{L_1(\rho_X)} \le \|f\|_{L_2(\rho_X)}.$$

We impose different assumptions on the unknown function f_ρ in Sections 2 and 3: in Section 2 we assume $f_\rho \in W$ and in Section 3 we assume $f_\rho\mu \in W$. It is clear that if $\mu(x)$ is a nice smooth function then for many smoothness classes W we have $f_\rho \in W \Rightarrow f_\mu \in W$. However, if μ is a "rough" function then the assumption $f_\rho \in W$ may be more appropriate.

Let us now discuss one more important issue. First, we remind the general scheme that we follow in constructing an estimator f_z. We begin with a function class W. Then, utilizing the *optimal method* approach we look for an estimator that provides good estimation for the class W. In some examples considered in Section 2 we choose a hypothesis space $\mathcal{H}$ where f_z comes from depending on the class W. It is a weak point of the above approach. In many cases we do not know exactly the class W. However, we may know a collection $\mathcal{W}$ of classes where our unknown class W belongs. Say, if we are thinking about W in terms of Sobolev smoothness classes we may take as $\mathcal{W}$ the collection of all Sobolev classes with smoothness from a certain range. We now modify the optimal method setting to the *universal method* setting. In this setting a collection $\mathcal{W}$ of classes is given and we need to find a procedure for constructing an estimator f_z in such a way that if $f \in W \in \mathcal{W}$ then $\|f - f_z\|$ is close to the optimal error for the class W. In approximation theory this approach is known under the name of universal method (see [T1–T4]). We would like to build a universal estimator f_z for a given collection $\mathcal{W}$ of classes. In Sections 2 and 3 we address this issue. We use different ideas in constructing universal estimators. In particular, the well known in nonlinear approximation theory and statistics the thresholding algorithm provides such a universal estimator for a collection of classes with finite smoothness.

By C and c we denote absolute positive constants and by $C(\cdot)$, $c(\cdot)$, and $A_0(\cdot)$ we denote positive constants that are determined by their arguments. We often have error estimates of the form $(\ln m/m)^\alpha$ that hold for $m \geq 2$. We could write these estimates in the form, say, $(\ln(m+1)/m)^\alpha$ to make them valid for all $m \in \mathbb{N}$. However, we use the first variant throughout the paper for the following two reasons: simpler notations, we are looking for the asymptotic behavior of the error.

2. Estimating f_ρ

Let ρ be a Borel probability measure on $Z = X \times Y$. If ξ is a random variable (a real valued function on a probability space Z) then denote

$$E(\xi) := \int_Z \xi d\rho; \quad \sigma^2(\xi) := \int_Z (\xi - E(\xi))^2 d\rho.$$

The following proposition gives a relation between $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ and $\|f - f_\rho\|_{L_2(\rho_X)}$.

Proposition 2.1 [CS]. *For every $f : X \to Y$*

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \int_X (f(x) - f_\rho(x))^2 d\rho_X.$$

We define the *empirical error* of f:

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

Let $f : X \to Y$. The *defect function* of f is

$$L_z(f) := L_{z,\rho}(f) := \mathcal{E}(f) - \mathcal{E}_z(f); \quad z = (z_1, \dots, z_m), \quad z_i = (x_i, y_i).$$

Theorem A [CS]. *Let $M > 0$ and $f : X \to Y$ be such that $|f(x) - y| \leq M$ a.e. Then, for all $\epsilon > 0$*

$$\mathrm{Prob}_{z\in Z^m}\{|L_z(f)| \leq \epsilon\} \geq 1 - 2\exp(-\frac{m\epsilon^2}{2(\sigma^2 + M^2\epsilon/3)}), \tag{2.1}$$

where $\sigma^2 := \sigma^2((f(x) - y)^2)$.

This theorem is a direct corollary of the Bernstein inequality: If $|\xi(z) - E(\xi)| \leq M$ a.e. Then for any $\epsilon > 0$

$$\mathrm{Prob}_{z\in Z^m}\{|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)| \geq \epsilon\} \leq 2\exp(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}).$$

Taking $\xi(z) := (f(x)-y)^2$ and noting that $E(\xi) = \mathcal{E}(f)$ we get (2.1).

Let X be a compact subset of $\mathbb{R}^d$. Denote $\mathcal{C}(X)$ the space of functions continuous on X with the norm

$$\|f\|_\infty := \sup_{x\in X} |f(x)|.$$

The paper [CS] indicates importance of a characteristic of a class W closely related to the concept of entropy numbers. For a compact subset W of a Banach space B we define the entropy numbers as follows

$$\epsilon_n(W,B) := \inf\{\epsilon : \exists f_1,\dots,f_{2^n} \in W : W \subset \cup_{j=1}^{2^n}(f_j + \epsilon U(B))\}$$

where $U(B)$ is the unit ball of Banach space B. We denote $N(W,\epsilon,B)$ the covering number that is the minimal number of balls of radius ϵ needed for covering W. In this paper in the most cases we take as a Banach space B the space $\mathcal{C} := \mathcal{C}(X)$ of continuous functions on a compact $X \subset \mathbb{R}^d$. We use the abbreviated notations

$$N(W,\epsilon) := N(W,\epsilon,\mathcal{C}); \quad \epsilon_n(W) := \epsilon_n(W,\mathcal{C}).$$

Theorem B [CS]. *Let W be a compact subset of $\mathcal{C}(X)$. Assume that for all $f \in W$, $f : X \to Y$ is such that $|f(x)-y| \le M$ a.e. Then, for all $\epsilon > 0$*

$$\text{Prob}_{z\in Z^m}\{\sup_{f\in W}|L_z(f)| \le \epsilon\} \ge 1 - N(W,\epsilon/(8M))2\exp(-\frac{m\epsilon^2}{2(\sigma^2+M^2\epsilon/3)}). \tag{2.2}$$

Here $\sigma^2 := \sigma^2(W) := \sup_{f\in W}\sigma^2((f(x)-y)^2)$.

We will give a proof of this theorem for completeness. We use the following simple relation.

Proposition 2.2 [CS]. *If $|f_j(x)-y| \le M$ a.e. for $j = 1,2$, then*

$$|L_z(f_1) - L_z(f_2)| \le 4M\|f_1 - f_2\|_\infty.$$

In the proof of this proposition we use

$$|(f_1(x)-y)^2 - (f_2(x)-y)^2| \le 2M\|f_1-f_2\|_\infty.$$

Proof of Theorem B. Let $f_1,\dots,f_N$ be the $\epsilon/(8M)$-net of W, $N := N(W,\epsilon/(8M))$. Then for any $f \in W$ there is an f_j such that $\|f - f_j\|_\infty \le \epsilon/(8M)$ and by Proposition 2.2

$$|L_z(f) - L_z(f_j)| \le \epsilon/2.$$

Therefore, $|L_z(f)| \ge \epsilon$ implies that there is a $j \in [1,N]$ such that $|L_z(f_j)| \ge \epsilon/2$, and

$$\text{Prob}_{z\in Z^m}\{\sup_{f\in W}|L_z(f)| \ge \epsilon\} \le \sum_{j=1}^N \text{Prob}_{z\in Z^m}\{|L_z(f_j)| \ge \epsilon/2\} \le 2N\exp(-\frac{m\epsilon^2}{8(\sigma^2+M^2\epsilon/6)}).$$

Denote by $f_{\mathcal{H}}$ a function from $\mathcal{H}$ that minimizes the error $\mathcal{E}(f)$:

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}}\mathcal{E}(f).$$

Theorem C [CS]. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \le M$ a.e. Then, for all $\epsilon > 0$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \le \epsilon\} \ge 1 - N(\mathcal{H}, \epsilon/(8M))2\exp(-\frac{m\epsilon^2}{8(4\sigma^2 + M^2\epsilon/3)}).$$

Here $\sigma^2 := \sigma^2(\mathcal{H}) := \sup_{f\in\mathcal{H}} \sigma^2((f(x) - y)^2)$.

This theorem follows from Theorem B and the chain of inequalities

$$0 \le \mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}_z(f_{z,\mathcal{H}}) + \mathcal{E}_z(f_{z,\mathcal{H}}) - \mathcal{E}_z(f_{\mathcal{H}}) + \mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}})$$

$$\le \mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}_z(f_{z,\mathcal{H}}) + \mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}).$$

Assume W is such that

$$\epsilon_n(W) \le C_1 n^{-r}, \quad n = 1, 2, \dots, \quad W \subset C_1 U(\mathcal{C}). \tag{2.3}$$

Then

$$N(W, \epsilon) \le 2^{(C_2/\epsilon)^{1/r}}. \tag{2.4}$$

Substituting this in Theorem C and optimizing over ϵ we get for $\epsilon = Am^{-\frac{r}{1+2r}}$, $A \ge A_0(M, r)$

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \le Am^{-\frac{r}{1+2r}}\} \ge 1 - \exp(-c(M)A^2 m^{\frac{1}{1+2r}}).$$

We have proved the following theorem that is essentially contained in [CS].

Theorem 2.1. *Assume that W is such that*

$$\epsilon_n(W) \le C_1 n^{-r}, \quad n = 1, 2, \dots, \quad W \subset C_1 U(\mathcal{C}).$$

Then for $A \ge A_0(M, r)$

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \le Am^{-\frac{r}{1+2r}}\} \ge 1 - \exp(-c(M)A^2 m^{\frac{1}{1+2r}}). \tag{2.5}$$

We will now impose some extra restictions on W and will get in return better estimates for $\mathcal{E}(f_z) - \mathcal{E}(f_W)$. We begin with the one from [CS] (see Theorem C* and Remark 13).

Theorem C* [CS]. *Suppose that either $\mathcal{H}$ is a compact and convex subset of $\mathcal{C}(X)$ or $\mathcal{H}$ is a compact subset of $\mathcal{C}(X)$ and $f_\rho \in \mathcal{H}$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \le M$ a.e. Then, for all $\epsilon > 0$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \le \epsilon\} \ge 1 - N(\mathcal{H}, \epsilon/(24M))2\exp(-\frac{m\epsilon}{288M^2}).$$

We will need the following theorem in a style of Theorem C*.

Theorem D. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \le M$ a.e. Then, for all $\epsilon > 0$*

$$\mathrm{Prob}_{z \in Z^m}\{\mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \le \epsilon\} \ge 1 - N(\mathcal{H}, \epsilon/(24M))2\exp(-\frac{m\epsilon}{C(M,K)})$$

under assumption $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le K\epsilon$.

Proof. The proof is similar to the proof of Theorem C* from [CS]. We will only point out the difference of the proofs. In the proof of Theorem C* the convexity assumption is used to prove the following lemma.

Lemma 2.1 [CS]. *Let $\mathcal{H}$ be a convex subset of $\mathcal{C}(X)$ such that $f_{\mathcal{H}}$ exists. Then for all $f \in \mathcal{H}$*

$$\|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

We use the assumption $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le K\epsilon$ instead of convexity and we use the following lemma instead of Lemma 2.1.

Lemma 2.2. *For any f we have*

$$\|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le 2(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + 2\|f_{\mathcal{H}} - f_\rho\|^2_{L_2(\rho_X)}).$$

Proof. We have

$$\|f - f_{\mathcal{H}}\|_{L_2(\rho_X)} \le \|f - f_\rho\|_{L_2(\rho_X)} + \|f_\rho - f_{\mathcal{H}}\|_{L_2(\rho_X)}.$$

Next,

$$\|f - f_\rho\|^2_{L_2(\rho_X)} = \mathcal{E}(f) - \mathcal{E}(f_\rho) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho).$$

Combining the above two relations we get

$$\|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le 2(\|f - f_\rho\|^2_{L_2(\rho_X)} + \|f_{\mathcal{H}} - f_\rho\|^2_{L_2(\rho_X)})$$

$$\le 2(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + 2\|f_{\mathcal{H}} - f_\rho\|^2_{L_2(\rho_X)}).$$

Thus instead of Lemma 2.1 we use the inequality

$$\|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le 2(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + 2K\epsilon)$$

in the proof of Theorem D.

The following theorem is essentially contained in [CS].

Theorem 2.2. *Assume that W satisfies (2.3). Suppose that either W is convex or $f_\rho \in W$. Then for $A \ge A_0(M,r)$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \le Am^{-\frac{r}{1+r}}\} \ge 1 - \exp(-c(M)Am^{\frac{1}{1+r}}).$$

The proof of this theorem repeats the proof of Theorem 2.1 with the following changes: we set $\epsilon = Am^{-\frac{r}{1+r}}$ and use Theorem C* to estimate $\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W)$.

We continue to consider classes satisfying a stronger assumption than (2.3). We first use the concept of the Kolmogorov width to impose an extra condition on W. Let us assume that W satisfies the following estimates for the Kolmogorov widths

$$d_n(W,\mathcal{C}) \le C_1 n^{-r}, \quad n = 1,2,\dots; \quad W \subset C_1 U(\mathcal{C}). \tag{2.6}$$

Then by Carl's [C] inequality

$$\epsilon_n(W) \le C_2 n^{-r}, \quad n = 1,2,\dots.$$

Therefore, for this class we have the estimate (2.5) as above in Theorem 2.1. We will prove a better estimate than (2.5).

Theorem 2.3. *Let $f_\rho \in W$ and let W satisfy (2.6). Then there exists an estimator f_z such that for $A \ge A_0(M,r)$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \le CA(\ln m/m)^{\frac{2r}{1+2r}}\} \ge 1 - \exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}). \tag{2.7}$$

Proof. Let a sequence $\{L_n\}$ be a sequence of optimal (near optimal) subspaces for W, $\dim L_n = n$. Then for any $f \in W$ there is a $\varphi_n \in L_n$ such that $\|f - \varphi_n\|_\infty \le C_1 n^{-r}$. It is clear that $\|\varphi_n\|_\infty \le 2\|f\|_\infty \le 2C_1$. We now consider in place of W from the above argument that gave Theorem 2.1 the set $V_n := 2C_1 U(\mathcal{C}) \cap L_n$. In other words we take as a hypothesis space the set V_n. Then it is well known that

$$N(V_n,\epsilon) \le (C_3/\epsilon)^n.$$

We construct an estimator for $f_\rho \in W$ by

$$f_{z,V_n} = \arg\min_{f\in V_n} \mathcal{E}_z(f).$$

Choosing $\epsilon = A(\ln m/m)^{\frac{2r}{1+2r}}$ and $n = [\epsilon^{-1/(2r)}] + 1$ we now estimate $\mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_\rho)$. Let $f^* \in V_n$ be such that

$$\|f_\rho - f^*\|_\infty \le C_1 n^{-r} \le CA^{1/2}(\ln m/m)^{\frac{r}{1+2r}}.$$

Then

$$\mathcal{E}(f^*) - \mathcal{E}(f_\rho) = \int_X (f^*(x) - f_\rho(x))^2 d\rho_X \le C^2 A(\ln m/m)^{\frac{2r}{1+2r}}. \tag{2.8}$$

We have

$$0 \le \mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_{z,V_n}) - \mathcal{E}(f^*) + \mathcal{E}(f^*) - \mathcal{E}(f_\rho).$$

Next,

$$\mathcal{E}(f_{z,V_n}) - \mathcal{E}(f^*) = \mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_{V_n}) + \mathcal{E}(f_{V_n}) - \mathcal{E}(f^*) \le \mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_{V_n}).$$

Taking into account the choice of $\epsilon = A(\ln m/m)^{\frac{2r}{1+2r}}$ and $n = [\epsilon^{-1/(2r)}] + 1$ we get from Theorem C* for $A > A_0(M,r)$

$$\text{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,V_n}) - \mathcal{E}(f^*) \le A(\ln m/m)^{\frac{2r}{1+2r}}\} \tag{2.9}$$

$$\ge 1 - \exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}).$$

Using (2.8) we obtain from here

$$\text{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_\rho) \le (1+C^2)A(\ln m/m)^{\frac{2r}{1+2r}}\} \tag{2.10}$$

$$\ge 1 - \exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}).$$

This completes the proof of Theorem 2.3.

We now proceed to imposing extra conditions on W in terms of nonlinear approximation. We begin with the definition of nonlinear Kolmogorov's (N,n)-width (see [T5]):

$$d_n(F,B,N) := \inf_{\mathcal{L}_N, \#\mathcal{L}_N \le N} \sup_{f\in F} \inf_{L\in\mathcal{L}_N} \inf_{g\in L} \|f-g\|_B,$$

where $\mathcal{L}_N$ is a set of at most N n-dimensional subspaces L. It is clear that

$$d_n(F,B,1) = d_n(F,B).$$

The new feature of $d_n(F,B,N)$ is that we allow to choose a subspace $L \in \mathcal{L}_N$ depending on $f \in F$. It is clear that the bigger N the more flexibility we have to approximate f. It turns out that from the point of view of our applications the following case

$$N \asymp n^{an},$$

where $a > 0$ is a fixed number, plays an important role.

Let us assume that W satisfies the following estimates for the nonlinear Kolmogorov widths

$$d_n(W,\mathcal{C},n^{an}) \le C_1 n^{-r}, \quad n = 1,2,\dots; \quad W \subset C_1 U(\mathcal{C}). \tag{2.11}$$

Then by [T5]

$$\epsilon_n(W)_\infty \le C_2(\ln n/n)^r, \quad n = 2,3,\dots.$$

For this class we have the estimate similar to (2.5) from above with m replaced by $m/\ln m$. It is clear that a class satisfying (2.11) is wider than a class satisfying $d_n(W,\mathcal{C}) \le C_1 n^{-r}$. We will prove an estimate for W satisfying (2.11) similar to (2.7).

Theorem 2.4. *Let $f_\rho \in W$ and let W satisfy (2.11). Then there exists an estimator f_z such that for $A \geq A_0(M, r, a)$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq CA(\ln m/m)^{\frac{2r}{1+2r}}\} \geq 1 - \exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}).$$

Proof. Denote for a given n a collection of $N_n := [n^{an}]$ optimal for W subspaces by $L_n^1, \dots, L_n^{N_n}$, $\dim L_n^j = n$. Then for any $f \in W$ there are a $j(f) \in [1, N_n]$ and a $\varphi_n \in L_n^{j(f)}$ such that $\|f - \varphi_n\|_\infty \leq C_1 n^{-r}$. It is clear that $\|\varphi_n\|_\infty \leq 2\|f\|_\infty \leq 2C_1$. We now consider the following set

$$U_n := \cup_{j=1}^{N_n} V_n^j, \quad V_n^j := 2C_1 U(\mathcal{C}) \cap L_n^j.$$

Then it is clear that

$$N(U_n, \epsilon) \leq N_n(C_3/\epsilon)^n.$$

We construct an estimator for $f_\rho \in W$ by

$$f_{z,U_n} = \arg\min_{f\in U_n} \mathcal{E}_z(f).$$

Choosing $\epsilon = A(\ln m/m)^{\frac{2r}{1+2r}}$ and $n = [\epsilon^{-1/(2r)}] + 1$ we now estimate $\mathcal{E}(f_{z,U_n}) - \mathcal{E}(f_\rho)$. Let $f^* \in U_n$ be such that

$$\|f_\rho - f^*\|_\infty \leq C_1 n^{-r} \leq C_1 A^{1/2}(\ln m/m)^{\frac{r}{1+2r}}.$$

Then

$$\mathcal{E}(f_{U_n}) - \mathcal{E}(f_\rho) \leq \mathcal{E}(f^*) - \mathcal{E}(f_\rho) = \int_X (f^*(x) - f_\rho(x))^2 d\rho_X \leq C_1^2 A(\ln m/m)^{\frac{2r}{1+2r}}. \tag{2.12}$$

We have

$$0 \leq \mathcal{E}(f_{z,U_n}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_{z,U_n}) - \mathcal{E}(f_{U_n}) + \mathcal{E}(f_{U_n}) - \mathcal{E}(f_\rho).$$

Taking into account that $f_{z,U_n} \in U_n$, $\mathcal{E}(f_{U_n}) - \mathcal{E}(f_\rho) \leq C_1^2\epsilon$,

$$N(V_n^j, \epsilon) \leq (C_3/\epsilon)^n, \quad j \in [1, N_n],$$

and the choice of $\epsilon = A(\ln m/m)^{\frac{2r}{1+2r}}$ and $n = [\epsilon^{-1/(2r)}] + 1$ we get from Theorem D

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,U_n}) - \mathcal{E}(f_{U_n}) \leq A(\ln m/m)^{\frac{2r}{1+2r}}\} \tag{2.13}$$

$$\geq 1 - \exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}), \quad A \geq A_0(M, r, a).$$

Using (2.12) we obtain from here

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,U_n}) - \mathcal{E}(f_\rho) \leq (1 + C_1^2)A(\ln m/m)^{\frac{2r}{1+2r}}\} \tag{2.14}$$

$$\geq 1-\exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}), \quad A \geq A_0(M,r,a).$$

This completes the proof of Theorem 2.4.

Let us discuss an example how Theorem 2.4 may be applied. We will consider n-term approximations with regard to a given system Ψ. Assume that the system $\Psi = \{\psi_j\}_{j=1}^\infty$ satisfies the condition:

(VP) There exist three positive constants A_i, $i=1,2,3$, and a sequence $\{n_k\}_{k=1}^\infty$, $n_{k+1} \leq A_1 n_k$, $k=1,2,\dots$ such that there is a sequence of the de la Vallée-Poussin type operators P_k with the properties

$$P_k(\psi_j) = \lambda_{k,j}\psi_j,$$

$$\lambda_{k,j} = 1 \quad \text{for} \quad j = 1,\dots,n_k; \qquad \lambda_{k,j} = 0 \quad \text{for} \quad j > A_2 n_k,$$

$$\|P_k\|_{\mathcal{C}\to\mathcal{C}} \leq A_3, \quad k = 1,2,\dots \quad .$$

Denote

$$\sigma_n(f,\Psi) := \inf_{k_1,\dots,k_n;c_1,\dots,c_n} \|f - \sum_{j=1}^n c_j \psi_{k_j}\|_\infty,$$

and

$$\sigma_n(W,\Psi) := \sup_{f\in W} \sigma_n(f,\Psi).$$

Theorem 2.5. *Let $f_\rho \in W$ and let W satisfy the following two conditions.*

$$\sigma_n(W,\Psi) \leq C_1 n^{-r}, \quad W \subset C_1 U(\mathcal{C}(X)).$$

$$E_n(W,\Psi) := \sup_{f\in W} \inf_{c_1,\dots,c_n} \|f - \sum_{j=1}^n c_j\psi_j\|_\infty \leq C_2 n^{-b},$$

where Ψ is the (VP)-system. Then there exists an estimator f_z such that for $A \geq A_0(M,r,b)$

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq CA(\ln m/m)^{\frac{2r}{1+2r}}\} \geq 1-\exp(-c(M)A(m(\ln m)^{2r})^{\frac{1}{1+2r}}).$$

Proof. Define

$$\Psi(n) := \{f : f = \sum_{j\in\Lambda} c_j\psi_j, |\Lambda| \leq n, \Lambda \subset [1,[n^{r/b}]+1], \|f\|_\infty \leq 2C_1\}.$$

As an estimator f_z we take $f_{z,\Psi(n)}$ with $n = [A^{-1/(2r)}(m/\ln m)^{\frac{1}{1+2r}}]+1$. For this n we consider the following family of n-dimensional subspaces:

$$X_\Lambda := \{f : f = \sum_{j\in\Lambda} c_j\psi_j, |\Lambda| = n\}, \quad \Lambda \subset [1,[n^{r/b}]+1].$$

Then the total number T of such subspaces satisfies

$$T \leq \binom{[n^{r/b}]+1}{n} \leq (n^{r/b}+1)^n \leq n^{(r/b+1)n}, \quad n \geq 2.$$

Thus it remains to apply Theorem 2.4 with $a = r/b + 1$.

We proceed to construction of universal estimators. Let us begin with a case where we impose conditions on the class W in a spirit of Kolmogorov's widths. Denote for a subspace L

$$d(W, L) := \sup_{f \in W} \inf_{g \in L} \|f - g\|_\infty.$$

Let $\mathcal{L} := \{L_n\}_{n=1}^\infty$ be a sequence of n-dimensional subspaces of $\mathcal{C}(X)$. Denote by $\mathcal{W}(\mathcal{L}, \alpha, \beta)$ a collection of classes $W^r(\mathcal{L})$, $r \in [\alpha, \beta]$, satisfying the following relations

$$d(W^r(\mathcal{L}), L_n) \leq C_1 n^{-r}, \quad n = 1, 2, \dots; \quad W^r(\mathcal{L}) \subset C_1(U(\mathcal{C}(X))).$$

Theorem 2.6. *For a given collection $\mathcal{W}(\mathcal{L}, \alpha, 1/2)$, $\alpha > 0$, there exists an estimator f_z such that if $f_\rho \in W^r(\mathcal{L})$, $r \in [\alpha, 1/2]$ then for $A \geq A_0(M, \alpha)$*

$$\text{Prob}_{z \in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq CA(\ln m/m)^{\frac{2r}{1+2r}}\} \geq 1 - m^{C_1(M)(C_2(M)-A)}.$$

Proof. We use the notations from the proof of Theorem 2.3. We define an estimator f_z by the formula

$$f_z := f_{z,V_k}$$

with

$$k = \arg\min_{1 \leq n \leq m} (\mathcal{E}_z(f_{z,V_n}) + An\ln m/m).$$

We will use the following result from [CS] (it is a direct corollary to Proposition 7 from [CS]).

Lemma 2.3. *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$. Then for all $\epsilon > 0$ with probability at least*

$$1 - N(\mathcal{H}, \frac{\epsilon}{24M}) \exp(-\frac{m\epsilon}{288M^2})$$

one has for all $f \in \mathcal{H}$

$$\mathcal{E}(f) \leq 2\mathcal{E}_z(f) + 2\epsilon - \mathcal{E}(f_\mathcal{H}) + 2(\mathcal{E}(f_\mathcal{H}) - \mathcal{E}_z(f_\mathcal{H})).$$

First of all we note that by Bernstein's inequality

$$\text{Prob}_{z \in Z^m}\{\mathcal{E}(f_\mathcal{H}) - \mathcal{E}_z(f_\mathcal{H}) \leq (A\ln m/m)^{1/2}\} \geq 1 - m^{C_3(M)(C_4(M)-A)}. \tag{2.15}$$

Applying Lemma 2.3 with $\mathcal{H} = V_n$, $\epsilon = An\ln m/m$, $f = f_{z,V_n}$ and using that $\mathcal{E}(f_{V_n}) \ge \mathcal{E}(f_\rho)$ we get for $n \in [1, m]$

$$\mathcal{E}(f_{z,V_n}) \le 2(\mathcal{E}_z(f_{z,V_n}) + An\ln m/m) - \mathcal{E}(f_\rho) + 2(A\ln m/m)^{1/2}$$

with probability at least $1 - m^{C_5(M)(C_6(M)-A)}$. Therefore,

$$\mathcal{E}(f_z) \le \min_{n\in[1,m]} 2(\mathcal{E}_z(f_{z,V_n}) + An\ln m/m) - \mathcal{E}(f_\rho) + 2(A\ln m/m)^{1/2}. \tag{2.16}$$

To estimate the right side we take $n = n(r) := [A^{-1/(2r)}(m/\ln m)^{1/(1+2r)}] + 1$. We have

$$\mathcal{E}_z(f_{z,V_{n(r)}}) \le \mathcal{E}_z(f_{V_{n(r)}}).$$

Similarly to (2.15) we get

$$\mathcal{E}_z(f_{V_{n(r)}}) \le \mathcal{E}(f_{V_{n(r)}}) + (A\ln m/m)^{1/2}$$

with probability $\ge 1 - m^{C_3(M)(C_4(M)-A)}$. Next, in the same way as we got (2.8) we obtain

$$\mathcal{E}(f_{V_{n(r)}}) \le \mathcal{E}(f_\rho) + C_1^2 A(\ln m/m)^{\frac{2r}{1+2r}}.$$

Substituting these estimates into (2.16) we get

$$\mathcal{E}(f_z) \le \mathcal{E}(f_\rho) + 2An(r)\ln m/m + 2C_1^2 A(\ln m/m)^{\frac{2r}{1+2r}} + 4(A\ln m/m)^{1/2} \tag{2.17}$$

with probability $\ge 1 - m^{C_1(M)(C_2(M)-A)}$.

This completes the proof of Theorem 2.6.

Some remarks. Let us compare the above new results with known results from [CS]. We will present only the estimates for

$$\epsilon(W, z) := \sup_{f_\rho\in W} \|f_\rho - f_z\|^2_{L_2(\rho_X)} \tag{2.18}$$

keeping in mind that the estimates hold with high probability in $z \in Z^m$.

Under the most general assumption on W in the form (2.3) Theorem 2.2 claims that there exists an estimator f_z with the error

$$\epsilon(W, z) \ll m^{-\frac{r}{1+r}}. \tag{2.19}$$

Now, if we impose an extra assumption on W in the form of decay of the Kolmogorov widths (2.6) then Theorem 2.3 gives

$$\epsilon(W, z) \ll (\ln m/m)^{\frac{2r}{1+2r}}. \tag{2.20}$$

This estimate is better than (2.19). In a particular case $W = W^r_2$ the Sobolev class on $[0,1]$ one can derive from (2.19) and from known estimates ([BS])

$$\epsilon_n(W^r_2, \mathcal{C}) \ll n^{-r}, \quad r > 1/2,$$

that there exists f_z such that

$$\epsilon(W^r_2, z) \ll m^{-\frac{r}{1+r}}, \quad r > 1/2. \tag{2.21}$$

It is known ([K]) that

$$d_n(W^r_2, \mathcal{C}) \ll n^{-r}, \quad r > 1/2. \tag{2.22}$$

Therefore, by Theorem 2.3 we get

$$\epsilon(W^r_2, z) \ll (\ln m/m)^{\frac{2r}{1+2r}}, \quad r > 1/2, \tag{2.23}$$

what is better than (2.21).

Theorems 2.4 and 2.5 show that the class W^r_2 can be replaced by a bigger class with (2.23) still holding. For instance, we may take $W = W^r_1$ - the Sobolev class defined in the L_1 norm. Then it is known that for $r > 1$ $W^r_1 \hookrightarrow W^{r-1}_\infty$ and for wavelet type system Ψ one has

$$\sigma_n(W^r_1, \Psi) \ll n^{-r}.$$

Therefore, we have by Theorem 2.5

$$\epsilon(W^r_1, z) \ll (\ln m/m)^{\frac{2r}{1+2r}}, \quad r > 1. \tag{2.24}$$

We also note that the estimate (2.22) is only an existence theorem. We do not know subspaces that provide approximation in (2.22). Therefore, Theorem 2.3 does not give a constructive estimator providing (2.23). However, we can use Theorem 2.5 to overcome this problem. For instance, in the periodic case it is known ([DT]) that

$$\sigma_n(W^r_2, \mathcal{T}) \ll n^{-r}, \quad r > 1/2,$$

where $\mathcal{T}$ is the trigonometric system. Using this result and the embedding $W^r_2 \hookrightarrow W^{r-1/2}_\infty$ we get by Theorem 2.5

$$\epsilon(W^r_2, z) \ll (\ln m/m)^{\frac{2r}{1+2r}}, \quad r > 1/2, \tag{2.25}$$

As a hypothesis space $\mathcal{H}$ one can take here the following set of n-term trigonometric polynomials

$$\mathcal{H} = \{f = \sum_{k\in\Lambda} c_k e^{ikx}, |\Lambda| \le n, \Lambda \subset [-n^{\frac{2r}{2r-1}}, n^{\frac{2r}{2r-1}}], \|f\|_\infty \le C\}$$

with $n \asymp (m/\ln m)^{\frac{1}{1+2r}}$.

3. Estimating f_μ

In this section we assume that ρ_X is an absolutely continuous measure with density $\mu(x)$: $d\rho_X = \mu dx$. We keep the notations from the previous section. Our major goal in this section is to estimate the function $f_\mu := f_\rho \mu$ instead of the function f_ρ. In this section we assume that $|y| \le M$. In the probability estimates we will use the following notations

$$w(m,A) := C_1 m^{C_2(M)(C_3(M)-A)}; \quad w(m,A;r) := C_1 m^{C_2(M)(C_3(M,r)-A)}$$

$$w(m,A;B,r) := C_1 m^{C_2(M,B)(C_3(M,B,r)-A)}$$

where C_1 an absolute constant, C_2, C_3 may depend on M and the indicated parameters.

3.1. Let $\{\psi_j\}$ be a uniformly bounded orthonormal basis for $L_2(X)$, $\|\psi_j\|_\infty \le B$. Assume that $f_\mu \in L_2(X)$. Then

$$f_\mu = \sum_{j=1}^{\infty} c_j \psi_j; \quad c_j := \int_X f_\mu \psi_j dx.$$

Denote

$$S_n(f_\mu) := \sum_{j=1}^{n} c_j \psi_j.$$

Consider

$$\hat{c}_j := \hat{c}_j(z) := \frac{1}{m} \sum_{i=1}^{m} y_i \psi_j(x_i).$$

Then

$$E(\hat{c}_j(z)) = \int_X f_\rho(x)\psi_j(x) d\rho_X = \int_X f_\mu(x)\psi_j(x)dx = c_j.$$

Therefore, by Bernstein's inequality applied to a random variable $y\psi_j(x)$ we obtain

$$\text{Prob}_{z\in Z^m}\{|\hat{c}_j(z) - c_j| \ge \eta\} \le 2\exp(-m\eta^2/C(M,B)). \tag{3.1}$$

Assume now that W is a class satisfying the following approximation property: for $f \in W$ one has

$$\|f - \sum_{j=1}^{n} c_j(f)\psi_j\|_2 \le C_1 n^{-r}, \quad c_j(f) := \int_X f\psi_j dx. \tag{3.2}$$

We define an estimator by the formula

$$f_{(z,n)} := \sum_{j=1}^{n} \hat{c}_j(z)\psi_j. \tag{3.3}$$

It is clear that

$$\|S_n(f_\mu) - f_{(z,n)}\|_2^2 = \sum_{j=1}^{n} |\hat{c}_j(z) - c_j|^2.$$

We get from (3.1)

$$\mathrm{Prob}_{z\in Z^m}\{|\hat{c}_j(z)-c_j|\le \eta, \quad j=1,\dots,n\}\ge 1-2n\exp(-m\eta^2/C(M,B)).$$

Using (3.2) we obtain from here

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu - f_{(z,n)}\|_2 \le n^{1/2}\eta + C_1 n^{-r}\}\ge 1-2n\exp(-m\eta^2/C(M,B)). \tag{3.4}$$

Choosing $\epsilon = (A\ln m/m)^{\frac{r}{1+2r}}$, $\eta=\epsilon n^{-1/2}$, $n=[\epsilon^{-1/r}]+1$ we get from (3.4)

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu - f_{(z,n)}\|_2 \le (1+C_1)(A\ln m/m)^{\frac{r}{1+2r}}\}\ge 1-w(m,A;B,r).$$

We have proved the following theorem.

Theorem 3.1. *Let Ψ be a uniformly bounded orthonormal basis, $\|\psi_j\|_\infty\le B$. Then for f_μ satisfying (3.2) the estimator $f_{(z,n)}$ defined by (3.3) with $n=[(m/(A\ln m))^{\frac{1}{1+2r}}]+1$ provides*

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu - f_{(z,n)}\|_2 \le (1+C_1)(A\ln m/m)^{\frac{r}{1+2r}}\}\ge 1-w(m,A;B,r).$$

3.2. In the previous subsection we considered the case of general uniformly bounded orthonormal basis. In this subsection we restrict ourselves to the case $\mu=1$ $(f_\rho=f_\mu)$ and also impose some additional (or other) assumptions on a basis Ψ and we obtain error estimates in the L_p-norm. We note that instead of assuming $\mu=1$ in the arguments that follow it is sufficient to assume that $\mu\le C$ with absolute constant C. Then we obtain the same results for f_μ instead of f_ρ. In order to illustrate new technique we consider a periodic case with a basis $\mathcal{T}$ the trigonometric system $\{e^{ikx}\}$. Let $\mathcal{V}_n(x)$ denote the de la Vallée-Poussin kernel. We define an estimator for f_ρ by the formula:

$$f_{z,VP} := \frac{1}{m}\sum_{i=1}^m y_i \frac{1}{\pi}\mathcal{V}_n(x-x_i). \tag{3.5}$$

Then for the random variable $\xi(y,u):=y\frac{1}{\pi}\mathcal{V}_n(x-u)$ we obtain

$$E(\xi)=\frac{1}{\pi}\int_0^{2\pi} f_\rho(u)\mathcal{V}_n(x-u)d\rho_X = \frac{1}{\pi}\int_0^{2\pi} f_\rho(u)\mathcal{V}_n(x-u)du =: V_n(f_\rho)(x).$$

It is well known that

$$\|V_n(f)\|_\infty \le C\|f\|_\infty.$$

Also, we have

$$E(\xi^2)\le C(M)n.$$

Let $x(l) = \pi l/(4n)$, $l = 1, \dots, 8n$. By Bernstein's inequality for each $l \in [1, 8n]$ we have

$$\mathrm{Prob}_{z\in Z^m}\{|V_n(f_\rho)(x(l)) - f_{z,VP}(x(l))| \ge \epsilon\} \le 2\exp(-\frac{m\epsilon^2}{C(M)n}).$$

Using the Marcinkiewicz-Zygmund [Z] theorem: for any trigonometric polynomial t of order N one has

$$\|t\|_\infty \le C_1 \max_{1\le k\le 4N} |t(k\pi/(2N))|$$

we obtain

$$\mathrm{Prob}_{z\in Z^m}\{\|V_n(f_\rho) - f_{z,VP}\|_\infty \le C_1\epsilon\} \ge 1 - nC_2\exp(-\frac{m\epsilon^2}{C(M)n}). \tag{3.6}$$

We define the class $W^r_p(\mathcal{T})$ as the set of f that satisfy the estimate:

$$\|f - V_n(f)\|_p \le C_3 n^{-r}, \quad 1 \le p \le \infty.$$

Assume that $f_\rho \in W^r_p(\mathcal{T})$. We specify $\epsilon = (A\ln m/m)^{\frac{r}{1+2r}}$, $n = [\epsilon^{-1/r}] + 1$. Then (3.6) implies

$$\mathrm{Prob}_{z\in Z^m}\{\|V_n(f_\rho) - f_{z,VP}\|_\infty \le C_1(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m, A)$$

and

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\rho - f_{z,VP}\|_p \le (C_1 + C_3)(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m, A).$$

We point out that in this subsection we have obtained the L_p estimates for $1 \le p \le \infty$. We conclude this subsection by formulating the result proved above as a theorem.

Theorem 3.2. *Assume $\mu = 1$ and $f_\rho \in W^r_p(\mathcal{T})$ with some $1 \le p \le \infty$. Then the estimator $f_{z,VP}$ defined by (3.5) with $n = [(m/(A\ln m))^{\frac{1}{1+2r}}] + 1$ provides*

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\rho - f_{z,VP}\|_p \le (C_1 + C_3)(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m, A).$$

We note that the estimator $f_{z,VP}$ from Theorem 3.2 does not depend on p and depends on r (the choice of n depends on r). We proceed to construction of an estimator that is universal for r. We denote

$$\mathcal{W}_p[\mathcal{T}] := \{W^r_p(\mathcal{T})\}.$$

Theorem 3.3. *For a given collection $\mathcal{W}_p[\mathcal{T}]$ there exists an estimator f_z such that if $f_\rho \in W^r_p(\mathcal{T})$ with some $r \le R$ then*

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\rho - f_z\|_p \le C(R)A^{1/2}(\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m, A).$$

Proof. We define

$$\mathcal{A}_0 := \mathcal{V}_1; \quad \mathcal{A}_s := \mathcal{V}_{2^s} - \mathcal{V}_{2^{s-1}}, \quad s = 1, 2, \dots; \quad A_s(f) := \mathcal{A}_s * f,$$

where $*$ means convolution. Using our assumption that $f_\rho \in W^r_p(\mathcal{T})$ we get for all s

$$\|A_s(f_\rho)\|_p \le K 2^{-rs} \tag{3.7}$$

with $K \le (1 + 2^R)C_3$. We consider the following estimators

$$f_{s,z} := \frac{1}{m} \sum_{i=1}^{m} y_i \mathcal{A}_s(x - x_i).$$

Similarly to (3.6) with $\epsilon = (A(2^s/m)\ln m)^{1/2}$ we get for all $s \in [0, \log m]$

$$\|A_s(f_\rho) - f_{s,z}\|_\infty \le C_1(A(2^s/m)\ln m)^{1/2} \tag{3.8}$$

with probability at least $1 - w(m, A)$. We now consider only those z that satisfy (3.8). We build an estimator f_z on the base of the sequence $\{\|f_{s,z}\|_p\}_{s=0}^{[\log m]}$. First, if

$$\|f_{s,z}\|_p \le (C_1 A^{1/2} + K)((2^s/m)\ln m)^{1/2}, \quad s = 0, \dots, [\log m], \tag{3.9}$$

then we set $f_z := 0$. We have in this case

$$\|f_\rho\|_p \le \sum_{s=0}^{\infty} \|A_s(f_\rho)\|_p. \tag{3.10}$$

Therefore, for z satisfying (3.8) and (3.9) we get from (3.8)–(3.10), (3.7) that

$$\|f_\rho\|_p \le C_1(R)A^{1/2} \sum_{s=0}^{\infty} \min(2^s \ln m/m)^{1/2}, 2^{-rs}) \le C_2(R)A^{1/2}(\ln m/m)^{\frac{r}{1+2r}}.$$

Second, if (3.9) is not satisfied then we let $l \in [0, \log m]$ be such that for $s \in (l, \log m]$

$$\|f_{s,z}\|_p \le (C_1 A^{1/2} + K)((2^s/m)\ln m)^{1/2}$$

and

$$\|f_{l,z}\|_p > (C_1 A^{1/2} + K)((2^l/m)\ln m)^{1/2}. \tag{3.11}$$

We set $n = 2^l$ and

$$f_z := \frac{1}{m} \sum_{i=1}^{m} y_i \mathcal{V}_n(x - x_i).$$

Then we get from (3.11) and (3.8)

$$\|A_l(f_\rho)\|_p \geq K((2^l/m)\ln m)^{1/2}.$$

Therefore, by (3.7) with $s = l$ we obtain

$$2^{l(1+2r)} \leq m/\ln m.$$

Let l_0 be such that

$$2^{(l_0-1)(1+2r)} \leq m/\ln m < 2^{l_0(1+2r)}.$$

Then for z satisfying (3.8) and not satisfying (3.9) we get

$$\|f_\rho - f_z\|_p \leq \|f_\rho - V_{2^{l_0}}(f_\rho)\|_p + \sum_{s=l+1}^{l_0} \|A_s(f_\rho)\|_p + \sum_{s=0}^{l} \|A_s(f_\rho) - f_{s,z}\|_p$$

$$\leq C_3 2^{-rl_0} + \sum_{s=l+1}^{l_0} (2C_1 A^{1/2} + K)((2^s/m)\ln m)^{1/2} + \sum_{s=0}^{l} C_1(A(2^s/m)\ln m)^{1/2}$$

$$\leq C(R)A^{1/2}(\ln m/m)^{\frac{r}{1+2r}}.$$

This completes the proof of Theorem 3.3.

We now point out that the above method that has been applied to the trigonometric system works also for more general systems. Let $\Psi = \{\psi_j\}_{j=1}^\infty$ be an orthonormal basis. First, we assume that Ψ is the (VP)-system for $\mathcal{C}$. Using notations from the definition of the (VP)-system we define

$$\mathcal{V}_{n_k}^{\Psi}(x,u) := \sum_j \lambda_{k,j}\psi_j(x)\psi_j(u).$$

Then

$$P_k(f) = \int_X f(u)\mathcal{V}_{n_k}^{\Psi}(x,u)du.$$

Second, we assume that for all k and $x, u \in X$

$$\text{(I)} \qquad |\mathcal{V}_{n_k}^{\Psi}(x,u)| \leq C' n_k, \quad \|\mathcal{V}_{n_k}^{\Psi}(x,\cdot)\|_2^2 \leq C' n_k.$$

Denote

$$\Psi(n) := \operatorname{span}\{\psi_1, \dots, \psi_n\}.$$

Third, we assume that for each n there exists a set of points $\xi^1, \dots, \xi^{N(n)} \in X$ such that $N(n) \leq n^c$ and for any $f \in \Psi(n)$

$$\text{(II)} \qquad \|f\|_\infty \leq C'' \max_i |f(\xi^i)|.$$

We define the class $W_p^r(\Psi)$ as the set of f satisfying

$$\|f - P_k(f)\|_p \leq C^1 n_k^{-r}, \quad 1 \leq p \leq \infty.$$

The following analog of Theorem 3.2 holds.

Theorem 3.2Ψ. *Let Ψ be an orthonormal basis. Suppose Ψ is the (VP)-system for $\mathcal{C}$ satisfying (I) and (II). Assume $f_\rho \in W^r_p(\Psi)$ with some $1 \le p \le \infty$. Let n_k be the smallest that satisfy $n_k \ge (m/(A\ln m))^{\frac{1}{1+2r}}$. Define*

$$f_z := \frac{1}{m}\sum_{i=1}^m y_i \mathcal{V}^\Psi_{n_k}(x,x_i).$$

Then

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\rho - f_z\|_p \le C(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m,A)$$

with constants that may depend on C', c, C'', C^1 and parameters A_1, A_2, A_3 from the definition of the (VP)-system.

The universality technique can also be extended to the bases Ψ from Theorem 3.2Ψ. We denote

$$\mathcal{W}_p[\Psi] := \{W^r_p(\Psi)\}.$$

Theorem 3.3Ψ. *Let Ψ be an orthonormal basis. Suppose Ψ is the (VP)-system for $\mathcal{C}$ satisfying (I) and (II). For a given collection $\mathcal{W}_p[\Psi]$ there exists an estimator f_z such that if $f_\rho \in W^r_p(\Psi)$ with some $r \le R$ then*

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\rho - f_z\|_p \le CA^{1/2}(\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m,A)$$

with constants that may depend on C', c, C'', C^1 and parameters A_1, A_2, A_3 from the definition of the (VP)-system.

The proof of this theorem goes along the lines of the proof of Theorem 3.3. We only explain the construction of f_z. Without loss of generality we may assume that the sequence $\{n_k\}$ from the definition of the (VP)-system satisfies the inequalities

$$A_1' n_k \le n_{k+1} \le A_1'' n_k, \quad A_1' > 1.$$

We define

$$\mathcal{A}^\Psi_0 := \mathcal{V}^\Psi_{n_1}; \quad \mathcal{A}^\Psi_s := \mathcal{V}^\Psi_{n_{s+1}} - \mathcal{V}^\Psi_{n_s}, \quad s = 1,2,\dots, \quad A^\Psi_s(f) := \int_X f(u)\mathcal{A}^\Psi_s(x,u)du.$$

Using our assumption that $f_\mu \in W^r_p(\Psi)$ we get

$$\|A^\Psi_s(f_\mu)\|_p \le K_1 n_s^{-r}.$$

We consider the following estimators

$$f^\Psi_{s,z} := \frac{1}{m}\sum_{i=1}^m y_i \mathcal{A}^\Psi_s(x,x_i).$$

Let l be such that for $s \in (l, \log m]$ (otherwise we set $f_z^\Psi := 0$)

$$\|f_{s,z}^\Psi\|_p \le (C'' A^{1/2} + K_1)((n_s/m)\ln m)^{1/2}$$

and

$$\|f_{l,z}^\Psi\|_p > (C'' A^{1/2} + K_1)((n_l/m)\ln m)^{1/2}.$$

We set

$$f_z^\Psi := \frac{1}{m}\sum_{i=1}^m y_i \mathcal{V}_{n_l}^\Psi(x, x_i).$$

3.3. In this subsection we extend the results of Section 3.1 to a wider range of function classes. We impose a weaker assumption than (3.2). It will be formulated in terms of nonlinear n-term approximations. Let Ψ be an orthonormal uniformly bounded basis for $L_2(X)$, $\|\psi_j\|_\infty \le B$. Denote

$$\sigma_n(f, \Psi)_2 := \inf_{k_1,\dots,k_n;c_1,\dots,c_n} \|f - \sum_{j=1}^n c_j \psi_{k_j}\|_2.$$

We will keep notations from subsection 3.1.

Theorem 3.4. *Let Ψ be an orthonormal uniformly bounded basis for $L_2(X)$, $\|\psi_j\|_\infty \le B$. Let $R \ge a > 0$ be given. We define the following estimator depending on a parameter A*

$$f_z := \sum_{j\in[1,m^{R/a}]:|\hat{c}_j(z)|\ge 2(A\ln m/m)^{1/2}} \hat{c}_j(z)\psi_j.$$

Assume f_μ satisfies

$$\sigma_n(f_\mu, \Psi)_2 \le C_1 n^{-r} \tag{3.12}$$

and

$$\|f_\mu - S_n(f_\mu)\|_2 \le C_2 n^{-a} \tag{3.13}$$

with $r \le R$. Then

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu - f_z\|_2 \le C(R)(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m, A; B, R/a).$$

Proof. We will be using the inequality (3.1) from subsection 3.1.

$$\mathrm{Prob}_{z\in Z^m}\{|\hat{c}_j(z) - c_j| \ge \eta\} \le 2\exp(-m\eta^2/C(M,B)). \tag{3.14}$$

Denote for some N that will be chosen later depending on m

$$\Lambda(z, \eta) := \{j \in [1, N] : |\hat{c}_j(z)| \ge 2\eta\},$$

and define an estimator for f_μ by

$$f_{z,\eta} := \sum_{j\in\Lambda(z,\eta)} \hat{c}_j(z)\psi_j.$$

Using (3.14) we obtain

$$\text{Prob}_{z\in Z^m}\{\max_{j\in[1,N]} |\hat{c}_j(z) - c_j| \le \eta\} \ge 1 - 2N\exp(-m\eta^2/C(M,B)). \tag{3.15}$$

Consider those z that satisfy $\max_{j\in[1,N]} |\hat{c}_j(z) - c_j| \le \eta$. For these z and $j \in \Lambda(z,\eta)$ we get $|c_j| \ge \eta$. Also, if j is such that $|c_j| \ge 3\eta$ then $j \in \Lambda(z,\eta)$. Moreover, we have

$$\|\sum_{j\in\Lambda(z,\eta)} (\hat{c}_j(z) - c_j)\psi_j\|_2 \le \eta|\Lambda(z,\eta)|^{1/2}. \tag{3.16}$$

It is not difficult to prove that for f_μ satisfying (3.12) one has

$$\#\{j : |c_j| \ge \eta\} \le C(R)\eta^{-\frac{2}{1+2r}}$$

and, therefore,

$$|\Lambda(z,\eta)| \le C(R)\eta^{-\frac{2}{1+2r}}. \tag{3.17}$$

Next,

$$\|\sum_{j\in[1,N]\setminus\Lambda(z,\eta)} c_j\psi_j\|_2 \le \|\sum_{j:|c_j|\le 3\eta} c_j\psi_j\|_2 \le C(R)\eta^{\frac{2r}{1+2r}}. \tag{3.18}$$

For all $r \le R$ we choose $N = [m^{R/a}]$ and $\eta = (A\ln m/m)^{1/2}$. Then combining (3.12), (3.13), (3.15), (3.16), (3.17), and (3.18) we obtain

$$\text{Prob}_{z\in Z^m}\{\|f_\mu - f_{z,\eta}\|_2 \le C(R)(A\ln m/m)^{\frac{r}{1+2r}}\} \ge 1 - w(m,A;B,R/a). \tag{3.19}$$

We stress here that the estimator f_z from Theorem 3.4 does not depend on r. In this sense the estimator f_z is universal. It adjusts automatically to the optimal smoothness of the class that contains f_μ.

3.4. This subsection is a natural continuation of the previous subsection. We assume in this subsection that $\{\psi_j\}$ is an orthonormal basis for $L_2(X)$ satisfying the condition $\|\psi_j\|_\infty \le Bj^{1/2}$, $j = 1, 2, \dots$ and impose an extra condition $\mu \le C$. Then instead of (3.14) we have by Bernstein's inequality

$$\text{Prob}_{z\in Z^m}\{|\hat{c}_j(z) - c_j| \ge \eta\} \le 2\exp(-\frac{m\eta^2}{C(M,B)(1+\eta j^{1/2})}). \tag{3.20}$$

We use the same notations and definitions as above in subsection 3.3. Instead of (3.15) we now have

$$\mathrm{Prob}_{z\in Z^m}\{\max_{j\in[1,N]}|\hat{c}_j(z)-c_j|\le\eta\}\ge 1-2N\exp(-\frac{m\eta^2}{C(M,B)(1+\eta N^{1/2})}). \tag{3.21}$$

Similarly to the above we get relation (3.16) and relations (3.17), (3.18) for a f_μ satisfying (3.12). We now set

$$\eta=(A\ln m/m)^{1/2},\quad N=[\eta^{-2}]+1.$$

Using (3.13) with $a=\frac{r}{1+2r}$ in the same way as we got (3.19) we obtain now a similar estimate

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu-f_{z,\eta}\|_2\le C(R)(A\ln m/m)^{\frac{r}{1+2r}}\}\ge 1-w(m,A;B,R). \tag{3.22}$$

We have proved the following theorem.

Theorem 3.5. *Let Ψ be an orthonormal basis for $L_2(X)$ satisfying the condition $\|\psi_j\|_\infty\le Bj^{1/2}$, $j=1,2,\dots$. Let $R>0$ be given. We define the estimator depending on a parameter A*

$$f_z:=\sum_{j\in[1,m/(A\ln m)]:|\hat{c}_j(z)|\ge 2(A\ln m/m)^{1/2}}\hat{c}_j(z)\psi_j.$$

Assume $\mu\le C$, f_μ satisfies (3.10) with $r\le R$ and satisfies (3.11) with $a=\frac{r}{1+2r}$. Then

$$\mathrm{Prob}_{z\in Z^m}\{\|f_\mu-f_z\|_2\le C(R)(A\ln m/m)^{\frac{r}{1+2r}}\}\ge 1-w(m,A;B,R).$$

In this case the estimator f_z is also universal.

We will make a remark how estimators for f_ρ and f_μ can be used. Let us denote here an estimator for f_ρ by f_z^ρ and an estimator for f_μ by f_z^μ. We mentioned in the Introduction that we can use f_z^μ to estimate $\int_S f_\rho d\rho_X$ by $\int_S f_z^\mu dx$ within the error $\|f_\mu-f_z^\mu\|_{L_1}$. We will now estimate the following integral: $\int_S f_\rho^2 d\rho_X$. We estimate it by $\int_S f_z^\rho f_z^\mu dx$. Suppose we have

$$\|f_\rho-f_z^\rho\|_{L_1(\rho)}\le\epsilon_1\quad\text{and}\quad\|f_\mu-f_z^\mu\|_{L_1}\le\epsilon_2.$$

Then

$$|f_\rho^2\mu-f_z^\rho f_z^\mu|\le|f_\rho\mu(f_\rho-f_z^\rho)|+|f_z^\rho||f_\rho\mu-f_z^\mu|.$$

Using $|f_\rho|\le M$, $|f_z^\rho|\le M$ we get

$$\|f_\rho^2\mu-f_z^\rho f_z^\mu\|_{L_1}\le M(\epsilon_1+\epsilon_2).$$

Therefore, for any $S\subset X$ the integral $\int_S f_z^\rho f_z^\mu dx$ gives the integral $\int_S f_\rho^2 d\rho_X$ within the error $M(\epsilon_1+\epsilon_2)$.

References

[BS] M.Sh. Birman and M.Z. Solomyak, *Estimates of singular numbers of integral operators*, Uspekhi Mat. Nauk **32** (1977), 17–84; English transl. in Russian Math. Surveys **32** (1977).

[C] B. Carl, *Entropy numbers, s-numbers, and eigenvalue problems*, J. Funct. Anal. **41** (1981), 290–306.

[CS] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bulletin of AMS, **39** (2001), 1–49.

[DT] R.A. DeVore and V.N. Temlyakov, *Nonlinear approximation by trigonometric sums*, J. Fourier Analysis and Applications **2** (1995), 29–48.

[K] B.S. Kashin, *Widths of certain finite-dimensional sets and classes of smooth functions*, Izv. Akad. Nauk SSSR, Ser.Mat. **41** (1977), 334–351; English transl. in Math. USSR IZV. **11** (1977).

[PS] T. Poggio and S. Smale, *The Mathematics of Learning: Dealing with Data*, manuscript (2003), 1–16.

[T1] V.N. Temlyakov, *Approximation by elements of a finite dimensional subspace of functions from various Sobolev or Nikol'skii spaces*, Matem. Zametki **43** (1988), 770–786; English transl. in Math. Notes **43** (1988), 444–454.

[T2] Temlyakov V.N., *On universal cubature formulas*, Dokl. Akad. Nauk SSSR **316** (1991), no. 1; English transl. in Soviet Math. Dokl. **43** (1991), 39–42.

[T3] V.N. Temlyakov, *Approximation of periodic functions*, Nova Science Publishes, Inc., New York, 1993.

[T4] V.N. Temlyakov, *Nonlinear Methods of Approximation*, Found. Comput. Math. **3** (2003), 33–107.

[T5] V.N. Temlyakov, *Nonlinear Kolmogorov's widths*, Matem. Zametki **63** (1998), 891–902.

[Z] A. Zygmund, *Trigonometric series*, University Press, Cambridge, 1959.