Industrial Mathematics Institute

2005:17

On universal estimators in learning theory

V.N. Temlyakov

IMI

Preprint Series

Department of Mathematics
University of South Carolina

ON UNIVERSAL ESTIMATORS IN LEARNING THEORY

V.N. Temlyakov

Abstract. This paper addresses a problem of constructing and analysing estimators for the regression problem in supervised learning. Recently, there was a big interest in studying universal estimators. Universal means that the estimator does not depend on an a priori assumption that the regression function f_ρ belongs to some class F from a collection of classes $\mathcal{F}$ and provides the estimation error for the f_ρ close to the optimal error for the class F. This paper is an illustration of how the general technique of construction of universal estimators, developed in the previous author's paper, can be applied in concrete situations. A setting of the problem studied in the paper has been motivated by a very recent paper by Smale and Zhou. The starting point for us is a given kernel $K(x,u)$ defined on $X \times \Omega$. On the base of this kernel we build an estimator that is universal for classes defined in terms of nonlinear approximations with regard to the system $\{K(\cdot,u)\}_{u\in\Omega}$. We apply the Relaxed Greedy Algorithm in construction of an estimator that is universal and easily implementable.

1. Introduction. Setting. Known results

This paper addresses a problem of constructing and analysing estimators for the regression problem in supervised learning. Recently, there was a big interest in studying universal estimators (see, for instance, [GKKW], [DKPT1,2], [KT], [BCDDT], [T1,2], [SZ], [KP], [CDD]). Universal means that the estimator does not depend on an a priori assumption that the regression function f_ρ belongs to some class F from a collection of classes $\mathcal{F}$ and provides the estimation error for the f_ρ close to the optimal error for the class F. The reader can find the rigorous definition of universally optimal estimators in [T2]. Two different general approaches to solving this problem (construction of a universal estimator) have emerged in recent works. In the first approach we begin with a collection $\mathcal{F} := \{F\}$ of classes of our interest and build a universal estimator for the collection $\mathcal{F}$. In the second approach we begin with a sequence $\{\mathcal{H}_n\}$ of hypothesis spaces where our estimators are supposed to come from. Then, we describe a collection $\mathcal{F}$ for which our estimators are universal. Clearly, these two approaches are closely related. In this paper we follow the lines of the second approach. A setting of the problem studied in the paper has been motivated by a very recent paper by Smale and Zhou [SZ]. They considered the following problem. Let a Mercer kernel K be given. This kernel generates the Reproducing Kernel Hilbert Space H_K with the norm $\|\cdot\|_K$. Next, they consider the regularized least square estimator: take a parameter $\lambda > 0$

and define for a given data $\mathbf{z} := (z_1, \dots, z_m)$, $z_i = (x_i, y_i)$,

$$f_{\lambda,\mathbf{z}} := \arg \min_{f\in H_K} \bigl(m^{-1}\sum_{i=1}^m (y_i - f(x_i))^2 + \lambda\|f\|_K^2\bigr). \tag{1.1}$$

It is known (see [CS,p.42]) that $f_{\lambda,\mathbf{z}}$ has the form

$$f_{\lambda,\mathbf{z}}(x) = \sum_{i=1}^m a_i K(x, x_i). \tag{1.2}$$

The representation (1.2) is a starting point for us. We will be looking for estimators of the form

$$\sum_{i=1}^m a_i K(x, u_i). \tag{1.3}$$

It is clear that an estimator of the form (1.3) can approximate a function f_ρ with accuracy δ with positive probability only in the case when the best approximation of f_ρ by functions of the form (1.3) is less than or equal to δ. We use this observation in forming a collection $\mathcal{F}$ of classes F that are well approximated by functions of the form (1.3). Namely, we consider the following classes. Let $K(x, u)$ be a continuous on $X \times \Omega$ function, where $X \subset \mathbb{R}^d$, $\Omega \subset \mathbb{R}^k$ are compact subsets. Define a system $\mathcal{K} := \{K(\cdot, u)\}_{u\in\Omega}$ and consider the restricted best m-term approximation of $f \in L_2(X, \mu)$, μ is a Borel measure on X, as

$$\sigma_{m,b}(f, \mathcal{K})_{L_2(X,\mu)} := \inf_{u_i\in\Omega, c_i:|c_1|+\dots+|c_m|\le b} \|f(x) - \sum_{i=1}^m c_i K(x, u_i)\|_{L_2(X,\mu)}.$$

Then for positive r, D we define

$$A^r(\mathcal{K}, D, b, \mu) := \{f \in L_2(X, \mu) : \sigma_{m,b}(f, \mathcal{K})_{L_2(X,\mu)} \le Dm^{-r}\}.$$

We point out that it is important that we define these classes using the restricted best m-term approximations $\sigma_{m,b}(f, \mathcal{K})_{L_2(X,\mu)}$. This allows us (under minor conditions on K) to apply the general technique of construction of universal estimators developed in [T2].

We now introduce the notations and formulate some theorems that we use in proofs of our results. Let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ be Borel sets, ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \to Y$ define *the error*

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

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 the conditional expectation

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

The function f_ρ is known in statistics as the *regression function* of ρ. It is clear that if $f_\rho \in L_2(\rho_X) := L_2(X,\rho_X)$ then it 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 $f_{\mathbf{z}}$, on the base of given data $\mathbf{z} = ((x_1,y_1),\dots,(x_m,y_m))$ that approximates f_ρ well with high probability. We assume that (x_i,y_i), $i = 1,\dots,m$ are independent and distributed accoding to ρ.

For a compact subset Θ of a Banach space B we define two variants of the entropy numbers as follows

$$e_a(\Theta,B) := \inf\{\epsilon : \exists f_1,\dots,f_{[a]} \in \Theta : \Theta \subset \cup_{j=1}^{[a]}(f_j + \epsilon U(B))\}, \quad a \ge 1,$$

$$\epsilon_n(\Theta,B) := e_{2^n}(\Theta,B), \quad n \in \mathbb{N},$$

where $U(B)$ is the closed unit ball of a Banach space B. We denote $N(\Theta,\epsilon,B)$ the covering number that is the minimal number of balls of radius ϵ with centers in Θ needed for covering Θ. We note that $N(\Theta,\epsilon_n(\Theta,B),B) \le 2^n$.

We proposed (see [DKPT2], [T2]) to study the following function that we call the *accuracy confidence function.* Let a set $\mathcal{M}$ of admissible measures ρ, and a sequence $\mathbb{E} := \{\mathbb{E}(m)\}_{m=1}^\infty$ of allowed classes $\mathbb{E}(m)$ of estimators be given. For $m \in \mathbb{N}$, $\eta > 0$ we define

$$\mathbf{AC}_m(\mathcal{M},\mathbb{E},\eta) := \inf_{E_m \in \mathbb{E}(m)} \sup_{\rho \in \mathcal{M}} \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \ge \eta\}$$

where E_m is an estimator that maps $\mathbf{z} \to f_{\mathbf{z}}$. For example, $\mathbb{E}(m)$ could be a class of all estimators, a class of linear estimators of the form

$$f_{\mathbf{z}} = \sum_{i=1}^m w_i(x_1,\dots,x_m,x)y_i,$$

or a specific estimator. In this paper we consider the case when $\mathbb{E}(m)$ is the set of all estimators, $m = 1,2,\dots$. We drop $\mathbb{E}$ from the notation and write $\mathbf{AC}_m(\mathcal{M},\eta)$.

By C and c we denote absolute positive constants and by $C(\cdot)$, $c(\cdot)$, and $A_0(\cdot)$ we denote constants that are determined by their arguments. For two nonnegative sequences $a = \{a_n\}_{n=1}^\infty$ and $b = \{b_n\}_{n=1}^\infty$ the relation (order inequality) $a_n \ll b_n$ means that there is a number $C(a,b)$ such that for all n we have $a_n \le C(a,b)\ b_n$; and the relation $a_n \asymp b_n$ means that $a_n \ll b_n$ and $b_n \ll a_n$.

We let μ be any Borel probability measure defined on X and let $\mathcal{M}(\Theta,\mu)$ denote the set of all ρ such that $\rho_X = \mu$, $|y| \le 1$, $f_\rho \in \Theta$. Denote by $\mathcal{C}(X)$ the space of continuous on X functions. The following theorem has been proved in [T2].

Theorem 1.1. *Let μ be a Borel probability measure on X. Assume $r > 0$ and Θ is a compact subset of $L_2(X,\mu)$ such that $\Theta \subset \frac{1}{4}U(\mathcal{C}(X))$ and*

$$\epsilon_n(\Theta,L_2(X,\mu)) \asymp n^{-r}. \tag{1.4}$$

Then there exist $\delta_0 > 0$ *and* $\eta_m^- \le \eta_m^+$, $\eta_m^- \asymp \eta_m^+ \asymp m^{-\frac{r}{1+2r}}$ *such that*

$$\mathbf{AC}_m(\mathcal{M}(\Theta,\mu),\eta) \ge \delta_0 \quad \textit{for} \quad \eta \le \eta_m^-$$

and

$$C_1 e^{-c_1(r)m\eta^2} \le \mathbf{AC}_m(\mathcal{M}(\Theta,\mu),\eta) \le e^{-c_2 m\eta^2}$$

for $\eta \ge \eta_m^+$.

Theorem 1.1 gives a very accurate description of the **AC**-function for classes Θ satisfying (1.4). This indicates that the behavior of the sequence $\{\epsilon_n(\Theta, L_2(X,\mu))\}$ determines the behavior of the sequence $\{\mathbf{AC}_m(\mathcal{M}(\Theta,\mu),\eta)\}$. Theorem 1.1 contains two components: the upper bounds and the lower bounds. In particular, the lower bounds show that the best accuracy we can achieve for classes satisfying (1.4) is not better than $\eta_m^- \asymp m^{-r/(1+2r)}$. We note that the lower bounds in Theorem 1.1 is a corollary (see [T1]) of the corresponding lower bounds from [DKPT2]. In Section 2 we prove some upper bounds by analysing the entropy numbers of a class of our interest. The lower bounds from Theorem 1.1 will serve as a benchmark for our method of proving the upper bounds.

We now proceed to results from [T2] on construction of universal (adaptive) estimators. We will use these results in Sections 3 and 4. Let a, β, be two positive numbers. Consider a collection $\mathcal{J}(a,\beta)$ of compacts J_n in $\mathcal{C}(X)$ satisfying

$$N(J_n,\epsilon,\mathcal{C}(X)) \le (a(1+1/\epsilon))^n n^{\beta n}, \quad n = 1,2,\dots. \tag{1.5}$$

Let us formulate a condition on measure ρ and a class $\mathcal{H}$ that we will often use:

$$\text{for all} \quad f \in \mathcal{H}, \quad \text{we have} \quad |f(x)-y| \le M \quad \text{a.e. with respect to} \quad \rho. \tag{1.6}$$

Clearly, (1.6) is satisfied if $|y| \le M/2$ and $|f(x)| \le M/2$, $f \in \mathcal{H}$.

The following two theorems from [T2] form a basis for construction of universal estimators. We use these theorems in Sections 3 and 4. Other examples of their use can be found in [T2]. We begin with the definition of our estimator. Let as above $\mathcal{J} := \mathcal{J}(a,\beta)$ be a collection of compacts J_n in $\mathcal{C}(X)$ satisfying (1.5).

We define

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

where

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

is the *empirical error (risk)* of f. This $f_{\mathbf{z},\mathcal{H}}$ is called the *empirical optimum* or the *least squares estimator*. We take a parameter $A \ge 1$ and consider the following estimator

$$f_{\mathbf{z}}^A := f_{\mathbf{z}}^A(\mathcal{J}) := f_{\mathbf{z},J_{n(\mathbf{z})}}$$

with

$$n(\mathbf{z}) := \arg\min_{1\le j\le m}\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},J_j}) + \frac{Aj\ln m}{m}\right).$$

Denote for a set L of a Banach space B

$$d(\Theta,L)_B := \sup_{f\in\Theta}\inf_{g\in L}\|f-g\|_B.$$

Theorem 1.2. *For $\mathcal{J} := \{J_n\}_{n=1}^\infty$ satisfying (1.5) and $M > 0$ there exists $A_0 := A_0(a,\beta,M)$ such that for any $A \geq A_0$ and any ρ such that ρ, J_n, $n = 1,2,\dots$ satisfy (1.6) we have*

$$\|f_{\mathbf{z}}^A - f_\rho\|^2_{L_2(\rho_X)} \leq \min_{1\leq j\leq m}(3d(f_\rho, J_j)^2_{L_2(\rho_X)} + \frac{4Aj\ln m}{m})$$

with probability $\geq 1 - m^{-c(M)A}$.

Theorem 1.3. *Let compacts $\{J_n\}$ satisfy (1.5) and $M > 0$ be given. There exists $A_0 := A_0(a,\beta,M) \geq 1$ such that for any $A \geq A_0$ and any ρ satisfying*

$$d(f_\rho, J_n)_{L_2(\rho_X)} \leq A^{1/2}n^{-r}, \quad n = 1,2,\dots,$$

and such that ρ, J_n, $n = 1,2,\dots,$ satisfy (1.6) we have for $\eta \geq A^{1/2}\big(\frac{\ln m}{m}\big)^{\frac{r}{1+2r}}$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)} \geq 4A^{1/2}\eta\} \leq Ce^{-c(M)m\eta^2}.$$

This paper is an illustration of how the general technique developed in [T2] can be applied in concrete situations. A discussion in Section 5 demonstrates that the above mentioned technique (based on Theorems 1.2 and 1.3) provides a powerful method of building estimators with good properties. We place the discussion section at the end of the paper for the reader's convenience. We present new results in Sections 2–4. In Section 4 we apply the Relaxed Greedy Algorithm in construction of an estimator that is universal and easily implementable.

2. Classes defined by integral operators

Let $\Omega \subset \mathbb{R}^k$ and $X \subset \mathbb{R}^d$ be compact subsets. Assume that the Borel probability measures ν and μ are defined on the sets Ω and X respectively. Suppose that $K(x,u)$ is a continuous function on $X \times \Omega$. We define the following integral operator

$$L_K(\varphi) := L_{K,\nu}(\varphi) := \int_\Omega K(x,u)\varphi(u)d\nu.$$

This operator is a compact operator that maps a Hilbert space $L_2(\Omega,\nu)$ into $L_2(X,\mu)$. For a positive number D we define a function class

$$W(K,D,\nu) := \{f : f = L_{K,\nu}(\varphi), \quad \|\varphi\|_{L_2(\Omega,\nu)} \leq D\}.$$

In this section we discuss the behavior of the $\mathbf{AC}_m(W(K,D,\nu),\mu)$. We will apply the general Theorem 1.1 for that purpose. Therefore, we need bounds of the sequence $\{\epsilon_n(W(K,D,\nu), L_2(X,\mu))\}$. There are known general results that give such bounds in terms of singular numbers of the operator L_K. E. Schmidt [S] gave an expansion (known as the Schmidt expansion)

$$K(x,u) = \sum_{j=1}^\infty s_j(L_K)\phi_j(x)\psi_j(u)$$

where $\{s_j(L_K)\}$ is a nonincreasing sequence of singular numbers of L_K, i.e. $s_j(L_K) := \lambda_j(L_K^*L_K)^{1/2}$, $\{\lambda_j(A)\}$ is a sequence of eigenvalues of an operator A, L_K^* is the adjoint operator to L_K. The two sequences $\{\phi_j(x)\}$ and $\{\psi_j(u)\}$ form orthonormal sequences of eigenfunctions of the operators $L_KL_K^*$ and $L_K^*L_K$ respectively.

Next, it is known that

$$d_n(W(K,D,\nu),L_2(X,\mu)) = s_{n+1}(L_K)D,$$

where $d_n(F,B)$ is the Kolmogorov width of an F in a Banach space B:

$$d_n(F,B) = \inf_{\{h_j\}_{j=1}^n} \sup_{f\in F} \inf_{\{c_j\}_{j=1}^n} \|f-\sum_{j=1}^n c_jh_j\|_B.$$

Finally, we use the following inequality due to Carl ([C]): for any $a>0$, we have

$$\max_{1\le k\le n} k^a\epsilon_k(F,B) \le C(a) \max_{1\le m\le n} m^a d_{m-1}(F,B).$$

In particular, the above argument implies that if

$$s_n(L_K) \le Cn^{-\alpha} \tag{2.0}$$

with some $\alpha>0$, then

$$\epsilon_n(W(K,D,\nu),L_2(X,\mu)) \le C(\alpha)n^{-\alpha}. \tag{2.1}$$

With (2.1) in hands we can apply the following general result from [T2].

Theorem 2.1. *Suppose ρ_X is fixed. Let $f_\rho \in \Theta$ and let ρ, Θ satisfy (1.6). Assume*

$$\epsilon_n(\Theta,L_2(\rho_X)) \le Dn^{-r}, \quad n=1,2,\dots, \quad \Theta \subset DU(L_2(\rho_X)).$$

Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \ge \epsilon_0$, $\epsilon_0 := C(M,D,r)m^{-\frac{r}{1+2r}}$, $m \ge 60(M/D)^2$, we have

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}-f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le \exp(-\frac{m\eta^2}{140M^2}).$$

Theorem 2.1 and (2.1) imply the following result.

Theorem 2.2. *Suppose ρ is such that $|y| \le M_1$. Assume that for a continuous kernel K we have*

$$s_n(L_{K,\nu}) \le D_1n^{-r}.$$

We set $M := M_1 + \|K\|_{C(X\times\Omega)}$. Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta\ge\epsilon_0$, $\epsilon_0 := C_1(M,D,D_1,r)m^{-\frac{r}{1+2r}}$, $m \ge C_2(M,D,D_1,r)$, we have for $f_\rho \in W(K,D,\nu)$, $\rho_X = \mu$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}-f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le \exp(-\frac{m\eta^2}{140M^2}).$$

Theorem 1.1 indicates that we cannot improve Theorem 2.2 if we use only (2.1). Also, in general, we cannot derive a better than (2.1) estimate from (2.0).

We now discuss some examples. Let $\Omega = X$ and $K(x,u)$ be a Mercer kernel (i.e. a function which is continuous, symmetric and for all finite sets $\{x_1,\dots,x_l\} \subset X$ the $l\times l$ matrix $\|K(x_i,x_j)\|_{i,j=1}^l$ is positive definite). Assume $\mu=\nu$ and use the following variant (see [CS,p.34]) of the classical Mercer's theorem.

Theorem 2.3. *For a Mercer kernel $K(x,u)$ we have*

$$K(x,u) = \sum_{l=1}^{\infty} \lambda_l(L_K)\phi_l(x)\phi_l(u), \tag{2.2}$$

where $\lambda_l(L_K)$ is the lth eigenvalue of L_K and ϕ_l is the corresponding eigenfunction. The convergence in (2.2) is absolute and

$$\sum_{l=1}^{\infty} \lambda_l(L_K) = \int_X K(x,x)d\nu \le C(K). \tag{2.3}$$

We note that in this case (L_K is self adjoint) we have $s_l(L_K) = \lambda_l(L_K)$. The bound (2.3) implies immediately that

$$\lambda_n(L_K) \le C(K)/n. \tag{2.4}$$

Let us consider a more general operator $L_K^r := L_{K,\nu}^r$, $r > 0$, than the operator L_K. We define for any $\varphi \in L_2(X,\nu)$

$$L_K^r(\varphi) := \sum_{l=1}^{\infty} \lambda_l(L_K)^r \langle \varphi, \phi_l\rangle \phi_l. \tag{2.6}$$

We note that $L_K^1 = L_K$. We associate the following class with the operator L_K^r

$$W^r(K,D,\nu) := \{f : f = L_K^r(\varphi), \quad \|\varphi\|_{L_2(X,\nu)} \le D\}.$$

Then it is known and easy to check that

$$d_n(W^r(K,D,\nu), L_2(X,\nu)) = \lambda_{n+1}(L_K)^r.$$

Therefore, by (2.1) and (2.4) we get

$$\epsilon_n(W^r(K,D,\nu), L_2(X,\nu)) \le C(r)Dn^{-r}. \tag{2.7}$$

Applying Theorem 2.1 we obtain the following result.

Theorem 2.4. *Suppose ρ is such that $|y| \le M_1$. Assume that K is the Mercer kernel. We set $M := M_1 + \|K\|_{C(X\times\Omega)}$. Let $r > 0$ and the Borel measure ν be fixed. Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \ge \epsilon_0$, $\epsilon_0 := C_1(M,D,K,r)m^{-\frac{r}{1+2r}}$, $m \ge C_2(M,D,K,r)$, we have for $f_\rho \in W^r(K,D,\nu)$, $\rho_X = \nu$*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le \exp(-\frac{m\eta^2}{140M^2}).$$

3. A universal estimator

Let X be a compact subset of $\mathbb{R}^d$ and let $B(X)$ be a Banach space of functions defined on X with the norm $\|f\|_{B(X)} := \sup_{x\in X}|f(x)|$. Consider a system $\mathcal{S} = \{g\}$ of functions from $B(X)$. We assume that the system $\mathcal{S}$ satisfies the following two conditions.

$$\mathcal{S} \subset C_1U(B(X)) \quad \text{that is} \quad \forall g \in \mathcal{S}, \quad |g(x)| \le C_1. \tag{3.1}$$

There exists a $\gamma > 0$ such that

$$e_n(\mathcal{S}, B(X)) \le C_2 n^{-\gamma}, \quad n = 1, 2, \dots. \tag{3.2}$$

We note that condition (3.2) implies (3.1).

We now describe a family of estimators based on the system $\mathcal{S}$. These estimators are defined depending on three parameters $q \ge 1$, $b > 0$, and $A > 0$. Let a set $\{g_l^n\}_{l=1}^{n^q}$ form an $e_{n^q}(\mathcal{S}, B(X))$-net of $\mathcal{S}$. We define the following compacts

$$F_n(q,b) := F_n(\mathcal{S},q,b) := \{f : \exists G \subset [1,n^q]\cap\mathbb{N}, |G| = n : f = \sum_{l\in G} c_l g_l^n, \quad \sum_{l\in G}|c_l| \le b\}.$$

For a parameter $A > 0$ we define

$$f_{\mathbf{z}}^A := f_{\mathbf{z}}^A(\mathcal{S}) := f_{\mathbf{z}}^{A,q,b}(\mathcal{S}) := f_{\mathbf{z},F_{n(\mathbf{z})}(q,b)}, \tag{3.3}$$

$$n(\mathbf{z}) := \arg\min_{1\le j\le m}(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},F_j(q,b)} + \frac{Aj\ln m}{m}).$$

We will prove in this section that estimators $f_{\mathbf{z}}^A$ perform well for the following classes

$$A^r(\mathcal{S}, D, b, \mu) := \{f \in L_2(X,\mu) : \sigma_{m,b}(f,\mathcal{S})_{L_2(X,\mu)} \le Dm^{-r}\},$$

where μ is a Borel probability measure on X and

$$\sigma_{m,b}(f,\mathcal{S})_{L_2(X,\mu)} := \inf_{g_i\in\mathcal{S}, c_i:|c_1|+\dots+|c_m|\le b} \|f - \sum_{i=1}^m c_i g_i\|_{L_2(X,\mu)}$$

is a restricted best m-term approximation of f with regard to the system $\mathcal{S}$.

Theorem 3.1. *Let positive numbers M_1, D, b, R be given. We set $M := M_1 + bC_1$. There exists an $A_0 := A_0(M, D, b, R, \gamma, C_2) \ge 1$ such that for the estimator $f_{\mathbf{z}}^A := f_{\mathbf{z}}^{A,q,b}(\mathcal{S})$ with $q := R/\gamma$ and $A \ge A_0$ we have the following error bounds. For any ρ satisfying $|y| \le M_1$, $f_\rho \in A^r(\mathcal{S}, D, b, \mu)$, $\rho_X = \mu$, with some $0 < r \le R$, μ - Borel probability measure, we have for $\eta \ge A^{1/2}\left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(X,\rho_X)} \ge 4A^{1/2}\eta\} \le Ce^{-c(M)m\eta^2}.$$

Proof. Our proof is based on Theorem 1.3. We will check that the conditions of Theorem 1.3 are satisfied. We specify the compacts $\{J_n\}$ from Theorem 1.3 in the following way $J_n := F_n(q,b)$. Then the estimator defined by (3.3) is exactly the estimator from Theorem 1.3. Thus, we need to check that the compacts $\{J_n\}$ satisfy the condition (1.5) with some a, β and check that the assumption $f_\rho \in A^r(\mathcal{S}, D, b, \mu)$, $\rho_X = \mu$, implies that

$$d(f_\rho, J_n)_{L_2(X,\mu)} \leq A^{1/2} n^{-r}, \quad n = 1, 2, \dots. \tag{3.4}$$

Also, we need to check that ρ, J_n satisfy (1.6). We begin with the latter property. By our assumption on ρ we have $|y| \leq M_1$ and by our assumption (3.1) on $\mathcal{S}$ we get from the definition of $F_n(q,b)$ that for $f \in F_n(q,b)$ $|f(x)| \leq bC_1$. Therefore, the condition (1.6) is satisfied with $M = M_1 + bC_1$.

We now proceed to the condition (1.5). For $G \subset [1, n^q] \cap \mathbb{N}$ with cardinality $|G| = n$ we define

$$F(G) := F(G,q,b) := \{f : f = \sum_{l \in G} c_l g_l^n, \quad \sum_{l \in G} |c_l| \leq b\}.$$

Then

$$F_n(q,b) = \cup_{G: G \subset [1,n^q] \cap \mathbb{N}, |G| = n} F(G). \tag{3.5}$$

Using $F(G) \subset bC_1 U(B(X))$ we obtain

$$N(F(G), \epsilon, B(X)) \leq (1 + 2bC_1/\epsilon)^n. \tag{3.6}$$

Then (3.5) and (3.6) imply

$$N(F_n(q,b), \epsilon, B(X)) \leq (1 + 2bC_1/\epsilon)^n \binom{[n^q]}{n} \leq n^{qn}(1 + 2bC_1/\epsilon)^n.$$

Therefore, condition (1.5) is satisfied with $a = \max(1, 2bC_1)$ and $\beta = q$.

It remains to check (3.4). Let $f_\rho \in A^r(\mathcal{S}, D, b, \mu)$, $\delta > 0$ and $\{c_i\}$, $\{g_i\} \subset \mathcal{S}$ be such that

$$\|f_\rho - \sum_{i=1}^n c_i g_i\|_{L_2(X,\mu)} \leq D n^{-r} + \delta, \quad \sum_{i=1}^n |c_i| \leq b.$$

For each $i \in [1,n]$ we find $g_{j(i)}^n$, $j(i) \in [1, n^q]$, such that

$$\|g_i - g_{j(i)}^n\|_{B(X)} \leq C_2 [n^q]^{-\gamma} \leq C(\gamma, C_2) n^{-q\gamma}.$$

Therefore,

$$d(f_\rho, F_n(q,b))_{L_2(X,\mu)} \leq \|f_\rho - \sum_{i=1}^n c_i g_{j(i)}^n\|_{L_2(X,\mu)} \leq D n^{-r} + \delta + bC(\gamma, C_2) n^{-q\gamma}. \tag{3.7}$$

Using $r \le R$ and $q = R/\gamma$ and taking into account that $\delta > 0$ is arbitrary, we obtain from (3.7) that

$$d(f_\rho, F_n(q,b))_{L_2(X,\mu)} \le (D + bC(\gamma, C_2))n^{-r}.$$

In order to satisfy (3.4) we choose A_0 such that $A_0 \ge (D + bC(\gamma, C_2))^2$.

To complete the proof it remains to apply Theorem 1.3.

Let us now discuss a particular example of a system $\mathcal{S}$ that satisfies (3.1) and (3.2). Consider a kernel $K(x,u)$ defined in the same way as in Section 2. In addition to our assumption that $K(x,u)$ is continuous on $X \times \Omega$ we assume that for some $\alpha > 0$ we have for any $u_1, u_2 \in \Omega$

$$\|K(\cdot,u_1) - K(\cdot,u_2)\|_{\mathcal{C}(X)} \le C_3\|u_1 - u_2\|^\alpha, \tag{3.8}$$

where $\|u\|$ is the eucledian norm of $u \in \mathbb{R}^k$. It is clear that continuity of $K(x,u)$ on the compact $X \times \Omega$ implies that

$$\|K\|_{\mathcal{C}(X\times\Omega)} \le C_4. \tag{3.9}$$

Consider a system $\mathcal{S} := \mathcal{K} := \{K(\cdot,u)\}_{u\in\Omega}$. Let us check that the $\mathcal{K}$ satisfies (3.1) and (3.2). Obviously, (3.9) implies (3.1). We now show that (3.8) implies (3.2). The set Ω is a compact subset of $\mathbb{R}^k$ and, therefore, for any $n \in \mathbb{N}$ there exists a net $u_1^n, \dots, u_n^n$ such that for any $u \in \Omega$

$$d(u, \{u_1^n, \dots, u_n^n\}) \le C(k)n^{-1/k}.$$

Then, using (3.8) we obtain

$$e_n(\mathcal{K}, B(X)) \le d(\mathcal{K}, \{K(\cdot,u_1^n), \dots, K(\cdot,u_n^n)\})_{\mathcal{C}(X)} \le C(k, C_3)n^{-\alpha/k}.$$

Therefore, (3.2) follows with $\gamma = \alpha/k$.

In this case ($\mathcal{S} = \mathcal{K}$), the class $A^r(\mathcal{K}, D, b, \mu)$ is the class of functions from $L_2(X,\mu)$ that can be approximated within an error Dn^{-r} by functions of the form

$$\sum_{i=1}^n c_i K(\cdot, u_i), \quad \sum_{i=1}^n |c_i| \le b. \tag{3.10}$$

Also, in this case, the estimator $f_{\mathbf{z}}^A$ takes the form (3.10) with $n \le m$. We formulate Theorem 3.1 in the case $\mathcal{S} = \mathcal{K}$.

Theorem 3.2. *Assume that the kernel $K(x,u)$ satisfies (3.8) and (3.9). Let positive numbers M_1, D, b, R be given. We set $M := M_1 + bC_3$. There exists an $A_0 := A_0(M, D, b, R, \alpha, C_3, C_4, k) \ge 1$ such that for the estimator $f_{\mathbf{z}}^A := f_{\mathbf{z}}^{A,q,b}(\mathcal{K})$ with $q := Rk/\alpha$ and $A \ge A_0$ we have the following error bounds. For any ρ satisfying $|y| \le M_1$, $f_\rho \in A^r(\mathcal{K}, D, b, \mu)$, $\rho_X = \mu$, with some $0 < r \le R$, μ - Borel probability measure, we have for $\eta \ge A^{1/2}\big(\frac{\ln m}{m}\big)^{\frac{r}{1+2r}}$*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(X,\rho_X)} \ge 4A^{1/2}\eta\} \le Ce^{-c(M)m\eta^2}.$$

4. Application of a greedy algorithm

It is well known (see, for instance, [H], [J], [LBW1,2], [CDD]) that greedy algorithms are useful in nonparametric statistics and learning theory. In this section we discuss an application of the Relaxed Greedy Algorithm (RGA) in finding an approximant for the estimator $f^A_{\mathbf{z}}$ considered in Section 3. There are different variants of the RGA (see [T3]). Here, we will discuss a variant of the RGA that is a generalization of the version of the RGA suggested by A. Barron ([B]). Let H be a real Hilbert space and let $\mathcal{G} := \{g\}$ be a system of elements $g \in H$ such that $\|g\| \le C_0$. Usually, in the theory of greedy algorithms we consider approximation with regard to a dictionary $\mathcal{D}$. One of the properties of a dictionary $\mathcal{D}$ is that $\overline{\operatorname{span}}\mathcal{D} = H$. In this section we do not assume that the system $\mathcal{G}$ is a dictionary. In particular, we do not assume that $\overline{\operatorname{span}}\mathcal{G} = H$. Denote $\mathcal{G}^\pm := \{\pm g, g \in \mathcal{G}\}$ the symmetrized system $\mathcal{G}$. Let $\theta > 0$.

RGA(θ) with regard to $\mathcal{G}$. *For $f \in H$ we define $f_0 := f$, $G_0 := G_0(f) := 0$. Then for each $n \ge 1$ we inductively define*

1) $\varphi_n \in \mathcal{G}^\pm$ is an element satisfying (we assume existence)

$$\langle f_{n-1}, \varphi_n \rangle = \max_{g \in \mathcal{G}^\pm} \langle f_{n-1}, g \rangle.$$

2)

$$G_n := G_n(f) := (1 - \frac{\theta}{n+\theta})G_{n-1} + \frac{\theta}{n+\theta}\varphi_n, \quad f_n := f - G_n.$$

Denote by $A_1(\mathcal{G})$ the closure in H of the convex hull of $\mathcal{G}^\pm$. Then for $f \in H$ there exists a unique element $f' \in A_1(\mathcal{G})$ such that

$$d(f, A_1(\mathcal{G}))_H = \|f - f'\| \le \|f - \phi\|, \quad \phi \in A_1(\mathcal{G}). \tag{4.1}$$

In analysis of the RGA(θ) we will use the following simple lemma (see [DT] for a variant of this lemma). Our analysis is similar to that of [DT] and [LBW1].

Lemma 4.1. *Let a sequence $\{a_n\}_{n=0}^\infty$ of nonnegative numbers satisfy the relations (with $\beta > 1$, $B > 0$)*

$$a_n \le \frac{n}{n+\beta}a_{n-1} + \frac{B}{(n+\beta)^2}, \quad n = 1, 2, \dots; \quad a_0 \le \frac{B}{(\beta-1)\beta}.$$

Then for all n

$$a_n \le \frac{B}{(\beta-1)(n+\beta)}.$$

Proof. Denoting $A := B/(\beta - 1)$ we obtain by induction

$$a_n \le \frac{A}{n-1+\beta}\frac{n}{n+\beta} + \frac{B}{(n+\beta)^2} = \frac{A}{n+\beta} - \frac{A(\beta-1)}{(n+\beta)(n-1+\beta)} + \frac{B}{(n+\beta)^2}.$$

Taking into account the inequality

$$\frac{A(\beta-1)}{(n+\beta)(n-1+\beta)} \ge \frac{A(\beta-1)}{(n+\beta)^2} = \frac{B}{(n+\beta)^2}$$

we complete the proof.

Theorem 4.1. *For $\theta > 1$ there exists a constant $C(\theta)$ such that for any $f \in H$ we have*

$$\|f_n\|^2 \leq d(f, A_1(\mathcal{G}))_H^2 + C(\theta)(\|f\| + C_0)^2 n^{-1}.$$

Proof. From the definition of G_n and f_n we get, denoting $\alpha := \frac{\theta}{n+\theta}$,

$$f_n = f - G_n = (1-\alpha)f_{n-1} + \alpha(f - \varphi_n)$$

and

$$\|f_n\|^2 = (1-\alpha)^2\|f_{n-1}\|^2 + 2\alpha(1-\alpha)\langle f_{n-1}, f - \varphi_n\rangle + \alpha^2\|f - \varphi_n\|^2. \tag{4.2}$$

It is known (see, for instance, [T4]) and easy to check that for any $h \in H$ one has

$$\sup_{g \in \mathcal{G}^\pm} \langle h, g\rangle = \sup_{\phi \in A_1(\mathcal{G})} \langle h, \phi\rangle. \tag{4.3}$$

Denote f' as above and set $f^* := f - f'$. Using (4.3) and the definition of φ_n we obtain from (4.2)

$$\|f_n\|^2 \leq (1-\alpha)^2\|f_{n-1}\|^2 + 2\alpha(1-\alpha)\langle f_{n-1}, f - f'\rangle + \alpha^2\|f - \varphi_n\|^2 =$$

$$(1-\alpha)(\|f_{n-1}\|^2 - \alpha\|f_{n-1}\|^2 + 2\alpha\langle f_{n-1}, f^*\rangle - \alpha\|f^*\|^2) + \alpha(1-\alpha)\|f^*\|^2 + \alpha^2\|f - \varphi_n\|^2.$$

This implies

$$\|f_n\|^2 - \|f^*\|^2 \leq (1-\alpha)(\|f_{n-1}\|^2 - \|f^*\|^2) + \alpha^2(\|f\| + C_0)^2.$$

Setting $a_n := \|f_n\|^2 - \|f^*\|^2$, $\beta := \theta$, and applying Lemma 4.1 we complete the proof.

Theorem 4.2. *For $\theta > 1/2$ there exists a constant $C := C(\theta, C_0)$ such that for any $f \in H$ we have*

$$\|f' - G_n(f)\|^2 \leq C/n,$$

Proof. If $f \in A_1(\mathcal{G})$ then the statement of Theorem 4.2 follows from known results ([B]). Assume that $d(f, A_1(\mathcal{G})) > 0$. Then the property (4.1) implies that for any $\phi \in A_1(\mathcal{G})$ we have

$$\langle f^*, \phi - f'\rangle \leq 0. \tag{4.4}$$

It follows from the definition of f_n that

$$f_n = (1 - \frac{\theta}{n+\theta})f_{n-1} + \frac{\theta}{n+\theta}(f - \varphi_n).$$

We set $f'_n := f_n - f^*$. Then, we get from the above representation

$$f'_n = (1 - \frac{\theta}{n+\theta})f'_{n-1} + \frac{\theta}{n+\theta}(f' - \varphi_n).$$

We note that $f'_n = f' - G_n(f)$. Let us estimate

$$\|f'_n\|^2 - \|f'_{n-1}\|^2 = \|f'_{n-1}\|^2((1 - \frac{\theta}{n+\theta})^2 - 1) + \tag{4.5}$$

$$\frac{2\theta}{n+\theta}(1 - \frac{\theta}{n+\theta})\langle f'_{n-1}, f' - \varphi_n\rangle + \frac{\theta^2}{(n+\theta)^2}\|f' - \varphi_n\|^2.$$

Next,

$$\langle f'_{n-1}, f' - \varphi_n\rangle = \langle f'_{n-1} + f^*, f' - \varphi_n\rangle - \langle f^*, f' - \varphi_n\rangle = \langle f_{n-1}, f' - \varphi_n\rangle + \langle f^*, \varphi_n - f'\rangle. \tag{4.6}$$

First, we prove that

$$\langle f_{n-1}, f' - \varphi_n\rangle \le 0. \tag{4.7}$$

It easily follows from $f' \in A_1(\mathcal{G})$ that

$$\langle f_{n-1}, f'\rangle \le \max_{g \in \mathcal{G}^\pm} \langle f_{n-1}, g\rangle. \tag{4.8}$$

By the definition of φ_n we get

$$\max_{g \in \mathcal{G}^\pm} \langle f_{n-1}, g\rangle = \langle f_{n-1}, \varphi_n\rangle. \tag{4.9}$$

Thus, (4.7) follows from (4.8) and (4.9).

Secondly, we note that (4.4) implies

$$\langle f^*, \varphi_n - f'\rangle \le 0. \tag{4.10}$$

Therefore, by (4.6), (4.7), and (4.10) we obtain

$$\langle f'_{n-1}, f' - \varphi_n\rangle \le 0. \tag{4.11}$$

Substitution of (4.11) in (4.5) gives

$$\|f'_n\|^2 - \|f'_{n-1}\|^2 \le \|f'_{n-1}\|^2(1 - \frac{2\theta}{n+\theta}) + \frac{\theta^2}{(n+\theta)^2}(\|f'_{n-1}\|^2 + \|f' - \varphi_n\|^2). \tag{4.12}$$

Using bounds $\|f'_{n-1}\| \le C_0$ and $\|f' - \varphi_n\| \le 2C_0$ we continue

$$\le \|f'_{n-1}\|^2(1 - \frac{2\theta}{n+\theta}) + 5C_0^2\theta^2/(n+\theta)^2.$$

We note that
$$1-\frac{2\theta}{n+\theta}<1-\frac{2\theta}{n+2\theta}.$$
We now apply Lemma 4.1 with $\beta=2\theta$ and get

$$\|f'_n\|^2\le C(\theta,C_0)/n. \tag{4.13}$$

This completes the proof.

We now discuss application of Theorem 4.1 in building an approximant for $f_{\mathbf{z},F_j(q,b)}$. Let, as in Section 3, the set $\{g^j_l\}_{l=1}^{j^q}$ form an $e_{j^q}(\mathcal{S},B(X))$-net of $\mathcal{S}$. Let $\mathbf{z}=(z_1,\dots,z_m)$, $z_i=(x_i,y_i)$, be given. Consider the following system of vectors in $\mathbb{R}^m$:
$$v^{j,l}:=(g^j_l(x_1),\dots,g^j_l(x_m)),\quad l\in[1,j^q].$$
We equip the $\mathbb{R}^m$ with the norm $\|v\|:=(m^{-1}\sum_{i=1}^m v_i^2)^{1/2}$. Then
$$\|v^{j,l}\|\le\|g^j_l\|_{B(X)}\le C_1.$$
Consider the following system in $H=\mathbb{R}^m$ with the defined above norm $\|\cdot\|$
$$\mathcal{G}:=\{v^{j,l}\}_{l=1}^{j^q}.$$
Finding the estimator
$$f_{\mathbf{z},F_j(\mathcal{S},q,b)}=\sum_{l\in\Lambda}c_lg^j_l,\quad \sum_{l\in\Lambda}|c_l|\le b,\quad |\Lambda|=j,\quad \Lambda\subset[1,j^q]\cap\mathbb{N},$$
is equivalent to finding best j-term approximant of $y\in\mathbb{R}^m$ from the $bA_1(\mathcal{G})$ in the space H. We apply the RGA(θ) with $\theta=2$ with respect to $\mathcal{G}$ to y/b and find, after j steps, an approximant
$$v^j:=\sum_{l\in\Lambda'}a_lv^{j,l},\quad \sum_{l\in\Lambda'}|a_l|\le 1,\quad |\Lambda'|=j,\quad \Lambda'\subset[1,j^q]\cap\mathbb{N},$$
such that
$$\|y/b-v^j\|^2\le d(y/b,A_1(\mathcal{G}))^2+Cj^{-1},\quad C=C(M_1,C_1).$$
We define an estimator
$$\hat f_{\mathbf{z}}:=\hat f_{\mathbf{z},F_j(q,b)}:=b\sum_{l\in\Lambda'}a_lg^j_l.$$
Then $\hat f_{\mathbf{z}}\in F_j(q,b)$ and
$$\mathcal{E}_{\mathbf{z}}(\hat f_{\mathbf{z},F_j(q,b)})\le\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},F_j(q,b)})+bCj^{-1}.$$

We denote $\delta := \{bCj^{-1}\}_{j=1}^m$ and define for $A \ge 1$

$$f^A_{\mathbf{z},\delta} := f^A_{\mathbf{z},\delta}(\mathcal{S}) := \hat{f}_{\mathbf{z},F_{n(\mathbf{z})}(q,b)}$$

with

$$n(\mathbf{z}) := \arg\min_{1\le j\le m}(\mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},F_j(q,b)}) + \frac{Aj\ln m}{m}).$$

By Remark 4.1 from [T2] we have for $A \ge A_0(M,b,\gamma,C_2)$

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \le \min_{1\le j\le m}(3d(f_\rho, F_j(q,b))^2 + \frac{4Aj\ln m}{m} + 2bCj^{-1/2}) \tag{4.14}$$

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

In particular, (4.14) means that the estimator $f^A_{\mathbf{z},\delta}$ is a universal estimator that provides the error

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \ll (\frac{\ln m}{m})^{\frac{2r}{1+2r}}$$

for f_ρ such that $\sigma_{m,b}(f_\rho,\mathcal{S})_{L_2(\rho_X)} \ll m^{-r}$, $r \le 1/2$. We note that the estimator $f^A_{\mathbf{z},\delta}$ is based on the greedy algorithm and it can easily be implemented.

5. Discussion

In this section we give a detailed comparison of our results with results from [SZ]. As we already mentioned in the Introduction the paper by Smale and Zhou [SZ] has served as a motivation for the study reported in this paper. First of all, we formulate the result from [SZ] that provides the estimation error for classes $W^r(K,D,\nu)$, defined in Section 2, in the case of Mercer kernel K.

Theorem 5.1. *Let $\mathbf{z}$ be randomly drawn according to ρ satisfying $|y| \le M$ almost surely. Assume that $f_\rho \in W^r(K,D,\nu)$ with some $r \in (0,1]$. We take $\delta \in (0,1)$ and set the regularization parameter*

$$\lambda := \log(4/\delta)(12\|K(x,x)\|^{1/2}_{\mathcal{C}(X)}M/D)^{2/(1+2r)}m^{-1/(1+2r)}, \quad r \in (1/2,1],$$

$$\lambda := 8\|K(x,x)\|_{\mathcal{C}(X)}\log(4/\delta)m^{-1/2}, \quad r \in (0,1/2].$$

Then, for $m \ge C(r)$ with confidence $1-\delta$,

$$\|f_{\lambda,\mathbf{z}} - f_\rho\|_{L_2(\rho_X)}$$

$$\le 2\log(4/\delta)(12\|K(x,x)\|^{1/2}_{\mathcal{C}(X)}M)^{2r/(1+2r)}D^{1/(1+2r)}m^{-r/(1+2r)}, \quad r \in (1/2,1], \tag{5.1}$$

$$\le \log(4/\delta)\big(8M + 8^r\|K(x,x)\|^r_{\mathcal{C}(X)}D\big)m^{-r/2}, \quad r \in (0,1/2]. \tag{5.2}$$

We compare this theorem with Theorem 2.4. The estimator $f_{\lambda,\mathbf{z}}$ in Theorem 5.1 does not depend on measure ν and does depend on $r \in (0,1]$. The estimator $f_{\mathbf{z}}$ from Theorem 2.4 may depend on both the measure ν and r. Thus, Theorem 5.1 gives the estimator $f_{\lambda,\mathbf{z}}$ that is universal for a collection $\{W^r(K,D,\nu)$, ν - Borel measure$\}$ with fixed $r \in (0,1]$. Theorem 2.4 is aimed at the optimal behavior of the estimator in the sense of error and confidence. For the minimal error ϵ_0 Theorem 2.4 gives $\epsilon_0 \asymp m^{-r/(1+2r)}$. Theorem 5.1 gives a similar bound for the error in the case $r \in [1/2,1]$ and the error of the order $m^{-r/2}$ in the case $r \in (0,1/2)$. In the range $r \in (0,1/2)$ we have $r/2 < r/(1+2r)$. Thus, Theorem 5.1 does not provide an optimal rate for the error of the estimator $f_{\lambda,\mathbf{z}}$ in the case $r \in (0,1/2)$. It would be interesting to understand if this effect is a result of analysis or reflects a property of the estimator $f_{\lambda,\mathbf{z}}$.

We now turn to the confidence bounds. Theorem 2.4 says that we have for $f_\rho \in W^r(K,D,\nu)$, $\rho_X = \nu$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \geq \eta\} \leq \exp\bigl(-\frac{m\eta^2}{140M^2}\bigr) \tag{5.3}$$

for $\eta \geq \epsilon_0$, $\epsilon_0 := C_1(M,D,K,r)m^{-\frac{r}{1+2r}}$, $m \geq C_2(M,D,K,r)$.

Let us compare the inequality (5.3) with the corresponding one from Theorem 5.1 in the case $r \in (1/2,1]$ when $f_{\lambda,\mathbf{z}}$ provides an optimal error rate. Rewriting the estimate (5.1) in the form of (5.3) gives

$$\rho^m\{\mathbf{z} : \|f_{\lambda,\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \geq \eta\} \leq 4\exp\bigl(-C(K,M,D,r)m^{r/(1+2r)}\eta\bigr). \tag{5.4}$$

The estimate (5.4) is not as good as the estimate (5.3). For instance, for $\eta \asymp m^{-r/(1+2r)}$ of a critical order, (5.3) gives $\exp(-c(M)m^{-1/(1+2r)})$ which is exponentially small and (5.4) gives $4\exp(-C(K,M,D,r))$ which does not approach 0 with $m \to \infty$. Also, we make the following important point. The estimator $f_{\mathbf{z}}$ does not depend on the target accuracy η and the estimator $f_{\lambda,\mathbf{z}}$ depends on it (λ depends on δ). It would be interesting to understand if the bounds like (5.4) is a price for universality with respect to measure ρ_X of the $f_{\lambda,\mathbf{z}}$ or is a result of analysis (or a specific feature of the estimator $f_{\lambda,\mathbf{z}}$).

As a conclusion of the above discussion we state that good features of the $f_{\lambda,\mathbf{z}}$ are its universality with respect to measure ρ_X and optimal error rate for $r \in [1/2,1]$. The drawbacks are: $f_{\lambda,\mathbf{z}}$ does not provide optimal error rate in the case $r \in (0,1/2)$ and it does not give good (optimal) bounds for the confidence.

We now proceed to a discussion of Theorem 3.2. This theorem gives the following bound for the estimator $f^A_{\mathbf{z}}$ for $\eta \geq A^{1/2}\bigl(\frac{\ln m}{m}\bigr)^{\frac{r}{1+2r}}$

$$\rho^m\{\mathbf{z} : \|f^A_{\mathbf{z}} - f_\rho\|_{L_2(X,\rho_X)} \geq 4A^{1/2}\eta\} \leq Ce^{-c(M)m\eta^2} \tag{5.5}$$

for $f_\rho \in A^r(\mathcal{K},D,b,\mu)$, $r \in (0,R]$, μ - Borel probability measure.

The estimator $f^A_{\mathbf{z}}$ does not depend on both r and μ. Therefore, the estimator $f^A_{\mathbf{z}}$ is universal for classes with different μ and with different smoothness $r \in (0,R]$. In this sense

$f_{\mathbf{z}}^A$ is better than the $f_{\lambda,\mathbf{z}}$. Next, for the minimal error of $f_{\mathbf{z}}^A$ we have $\asymp \big(\frac{\ln m}{m}\big)^{r/(1+2r)}$ which is very close to the optimal order of $m^{-r/(1+2r)}$. The estimator $f_{\mathbf{z}}^A$ provides this minimal error for all $r \in (0, R]$. Finally, (5.5) gives optimal bounds for the confidence. Therefore, the estimator $f_{\mathbf{z}}^A$ is very good with respect to the theoretical criteria: universality, minimal error bounds, confidence bounds. As we showed in Section 4, a variant of $f_{\mathbf{z}}^A$ can be easily implemented in practice using the Relaxed Greedy Algorithm. However, our analysis in this case works only for $r \le 1/2$. It would be very interesting to find a way of practical implementation of the $f_{\mathbf{z}}^A$ or its variant in the case $r \le R$ with big R.

We now proceed to a comparison of settings in [SZ] and in our paper. First, we study different function classes: the classes $W^r(K, D, \nu)$, defined in a standard for linear approximation theory way, are studied in [SZ]; The classes $A^r(\mathcal{K}, D, b, \mu)$, defined in a nonlinear approximation way, are studied in this paper. Second, we impose different assumptions on the kernel K: K is the Mercer kernel in [SZ]; K is any Hölder smooth in one variable kernel in this paper. Third, we use different general methods for construction of estimators: [SZ] builds $f_{\lambda,\mathbf{z}}$ as the regularized least squares estimator; we build $f_{\mathbf{z}}^A$ as the penalized least squares estimator. Clearly, it would be very interesting to make a bridge between methods from [SZ] and our methods. In particular, it is very interesting to understand a relation between classes $W^r(K, D, \nu)$ and $A^r(\mathcal{K}, D, b, \nu)$. Usually, nonlinear classes are bigger than their linear counterparts. We give here an example where one can see the above mentioned phenomenon.

Example. Let $X = \Omega = [0, 1]$ and let $K(x, u) = \chi(x - u)$ where $\chi(x) = 1$ if $x \ge 0$, $\chi(x) = 0$ if $x < 0$. Then the operator L_K acts as follows

$$f(x) = L_K(\varphi)(x) = \int_0^1 \chi(x-u)\varphi(u)d\nu = \int_0^x \varphi(u)d\nu, \quad x \in [0, 1]. \tag{5.6}$$

We begin with a simple observation about the variation $V(f)$ of function $f(x)$:

$$V(f) \le \int_0^1 |\varphi(u)|d\nu = \|\varphi\|_{L_1(\nu)} \le \|\varphi\|_{L_2(\nu)}.$$

Therefore, for any Borel probability measure ν we have $W(K, D, \nu) \subseteq BV(D)$, where $BV(D)$ is the class of functions of bounded variation with the bound D.

Consider now the class $A^1(\mathcal{K}, D, b, \mu)$. Denoting

$$f_+(x) := \int_0^x |\varphi(u)|d\nu, \quad f_-(x) := \int_0^x (|\varphi(u)| - \varphi(u))d\nu,$$

we obtain the classical representation $f = f_+ - f_-$ of the function f of bounded variation as a difference of two monotone functions. Clearly, $f_+(0) = f_-(0) = 0$ and $f_+(1) \le D$, $f_-(1) \le 2D$. It is easy to see that for $b \ge 2D$ we have

$$\sigma_{n,b}(f_+, \mathcal{K})_{B([0,1])} \le D/n.$$

Therefore, for $b \geq 6D$ we have

$$\sigma_{2n,b}(f, \mathcal{K})_{B([0,1])} \leq 3D/n.$$

This bound implies that

$$\sigma_{m,6D}(f, \mathcal{K})_{L_2(\mu)} \leq 12D/m, \quad m = 1, 2, \dots.$$

Thus, we have obtained the following embedding in the particular case $K(x,u) = \chi(x-u)$

$$W(K, D, \nu) \subseteq A^1(\mathcal{K}, 12D, 6D, \mu).$$

It is clear from the above argument that the class $W(K, D, \nu)$ is, in general, much smaller than the class $A^1(\mathcal{K}, 12D, 6D, \mu)$.

References

[B] Andrew R. Barron, *Universal approximation bounds for superposition of n sigmoidal functions*, IEEE Transactions on Information Theory **39** (1993), 930– 945.

[BCDDT] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov, *Universal algorithms for learning theory. Part I: piecewise constant functions*, Manuscript (2004), 1–24.

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

[CDD] A. Cohen, W. Dahmen, and R. DeVore, *Approximation and learning by greedy algorithms*, Manuscript (2005), 1–27.

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

[DKPT1] R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, *On Mathematical Methods of Learning*, IMI Preprints **10** (2004), 1–24.

[DKPT2] R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, *Mathematical methods for supervised learning*, IMI Preprints **22** (2004), 1–51.

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

[DT] R.A. DeVore and V.N. Temlyakov, *Some remarks on Greedy Algorithms*, Advances in Computational Mathematics **5** (1996), 173–187.

[H] P.J. Huber, *Projection Pursuit*, The Annals of Statistics **13** (1985), 435–475.

[J] L. Jones, *On a conjecture of Huber concerning the convergence of projection pursuit regression*, Annals of Stat. **15** (1987), 880–882.

[KP] G. Kerkyacharian and D. Picard, *Thresholding in Learning Theory*, Manuscript (2005), 1–21.

[KT] S. Konyagin and V. Temlyakov, *The Entropy in the Learning Theory. Error Estimates*, IMI Preprints **09** (2004), 1–25.

[LBW1] W.S. Lee, P.L. Bartlett, and R.C. Williamson, *Efficient agnostic learning of neural networks with bounded fan-in*, IEEE Transactions on Information Theory **42(6)** (1996), 2118–2132.

[LBW2] W.S. Lee, P.L. Bartlett, and R.C. Williamson, *The importance of convexity in learning with squared loss*, IEEE Transactions on Information Theory **44(5)** (1998), 1974–1980.

[S] E. Schmidt (1906-1907), *Zur Theorie der linearen und nichtlinearen Integralgleichungen. I*, Math. Annalen **63** (1906-1907), 433–476.

[SZ] S. Smale and D-X. Zhou, *Learning Theory Estimates via Integral Operators and Their Approximations*, Manuscript (2005), 1–23.

[T1] V.N. Temlyakov, *Optimal Estimators in Learning Theory*, IMI Preprints **23** (2004), 1–29.

[T2] V.N. Temlyakov, *Approximation in Learning Theory*, IMI Preprints **05** (2005), 1–42.

[T3] V.N. Temlyakov, *Nonlinear Methods of Approximation*, Found. Comput. Math. **3** (2003), 33–107.

[T4] V.N. Temlyakov, *Greedy algorithms in Banach spaces*, Advances in Comput. Math. **14** (2001), 277–292.