INDUSTRIAL
MATHEMATICS
INSTITUTE

2004:05

Some error estimates in learning theory

S. Konyagin and V.N. Temlyakov

IMI
Preprint Series

Department of Mathematics
University of South Carolina

SOME ERROR ESTIMATES IN LEARNING THEORY

S.V. Konyagin and V.N. Temlyakov

Abstract. We continue investigation of some problems in learning theory in the setting formulated by F. Cucker and S. Smale [CS]. The 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 well the regression function f_ρ of an unknown Borel probability measure ρ defined on $Z = X \times Y$. Following [CS] we consider a problem of approximate recovery of a projection f_W of an unknown regression function f_ρ onto a given class of functions W. It is known from [CS] and [DKPT] that the behavior of the entropy numbers $\epsilon_n(W)$ of W in the uniform norm plays an important role in the above problem. In this paper we obtain sharp (in the sense of order) estimates for the error between f_W and f_z for the classes W satisfying $\epsilon_n(W) \le Dn^{-r}$, $n = 1, 2, \dots$, $|f| \le D$, $f \in W$. We observe that the error estimates exhibit a saturation phenomenon for the range $r > 1/2$. We improve the error estimates by imposing one additional assumption on the relation between f_ρ and W, namely, we assume $f_\rho \in W$.

We discuss one more issue in the paper. We provide a method that calculates from the data z an approximate value of the average variance $\int_Z (y - f_\rho(x))^2 d\rho$ of the random variable y with controlled error estimate.

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], [DKPT]).

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*

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

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 if $f_\rho \in L_2(\rho_X)$ 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_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. 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 W in Banach space B is defined as follows

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

where $\inf_L$ is taken over all n-dimensional linear subspaces of B. In other words the Kolmogorov n-width gives the best possible error in approximating a compact set W by n-dimensional linear subspaces. So, first of all we need to choose a function class W (a hypothesis space $\mathcal{H}$ in [CS]) to work with. After selecting a class W we have the following two ways to go. The first one ([CS], [PS]) is based on the idea of studying approximation of a projection f_W of f_ρ onto W. In this case we do not assume that the regression function f_ρ comes from a specific (say, smoothness) class of functions. The second way ([CS], [PS], [DKPT]) is based on the assumption $f_\rho \in W$. For instance, we may assume that f_ρ has some smoothness. The next step is to find a method for constructing an estimator f_z that provides a good (optimal, near optimal in a certain sense) error $\|f_\rho - f_z\|$ for all $f_\rho \in W$ with high probability with respect to ρ. A problem of optimization is naturally broken into two parts: upper estimates and lower estimates. 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]) (*approximation space* [DKPT]) where an estimator f_z comes from.

The next question is how to build $f_z \in \mathcal{H}$. In this paper we discuss a standard in statistics method of *empirical risk minimization* 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 (risk)* of f. This $f_{z,\mathcal{H}}$ is called the *empirical optimum.* Section 2 contains a discussion of known results from [CS], [DKPT] and some new results. We obtain results in both directions mentioned above. Proofs of new results in Section 2 are based on the combination of the technique developed in [CS] and [DKPT] and new technique developed here.

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 $N(W,\epsilon)$ 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 note that for a fixed ρ_X all our results hold with $\mathcal{C}(X)$ replaced by $L_\infty(\rho_X)$. However, we formulate all results using the space $\mathcal{C}(X)$ because we want to have assumptions on W independent of ρ. Following [DKPT] we impose restrictions on a class W in the following two forms:

$$\epsilon_n(W) := \epsilon_n(W,\mathcal{C}) \le Dn^{-r}, \quad n = 1,2,\dots, \quad W \subset DU(\mathcal{C}), \tag{1.1}$$

or

$$d_n(W) := d_n(W,\mathcal{C}) \le Kn^{-r}, \quad n = 1,2,\dots, \quad W \subset KU(\mathcal{C}). \tag{1.2}$$

After building f_z we need to choose an appropriate norm $\|\cdot\|$ to measure the error $\|f_\rho - f_z\|$. In [CS] the quality of approximation is measured by $\mathcal{E}(f_z) - \mathcal{E}(f_\rho)$. It is easy to see that for any $f \in L_2(\rho_X)$

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L_2(\rho_X)}. \tag{1.3}$$

Thus the choice $\|\cdot\| = \|\cdot\|_{L_2(\rho_X)}$ seems natural. This norm has also been used in [DKPT] for measuring the error. One of important questions is to estimate the *defect function* $L_z(f) := \mathcal{E}(f) - \mathcal{E}_z(f)$ of $f \in W$. If ξ is a random variable (a real valued function on a probability space Z) then denote

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

For a single function f the following theorem from [CS] is a corollary of the probabilistic Bernstein inequality: if $|\xi(z) - E(\xi)| \le M$ a.e. then for any $\epsilon > 0$

$$\mathrm{Prob}_{z\in Z^m}\{|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)| \ge \epsilon\} \le 2\exp(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}). \tag{1.4}$$

Theorem 1.1 [CS]. *Let $M > 0$ and $f : X \to Y$ be such that $|f(x) - y| \le M$ a.e. Then, for all $\epsilon > 0$*

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

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

We will assume that ρ and W satisfy the following condition.

$$\text{For all} \quad f \in W, \quad f : X \to Y \quad \text{is such that} \quad |f(x) - y| \le M \quad \text{a.e.} \tag{1.5}$$

The following useful inequality has been obtained in [CS].

Theorem 1.2 [CS]. *Let W be a compact subset of $\mathcal{C}(X)$. Assume that ρ, W satisfy (1.5). Then, for all $\epsilon > 0$*

$$\text{Prob}_{z\in Z^m}\{\sup_{f\in W}|L_z(f)| \ge \epsilon\} \le N(W,\epsilon/(8M))2\exp(-\frac{m\epsilon^2}{2(\sigma^2+M^2\epsilon/3)}). \tag{1.6}$$

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

This theorem contains a factor $N(W,\epsilon/(8M))$ that may grow exponentially for classes W satisfying (1.1): $N(W,\epsilon) \le 2^{(D/\epsilon)^{1/r}+1}$. In Section 2 we prove a stronger (in a certain sense) estimate than (1.6) under assumption that W satisfies (1.1). For instance, in the case $r > 1/2$ Theorem 2.2 replaces $N(W,\epsilon/(8M))$ in an analogue of (1.6) by a constant $C(M,D,r)$ independent of ϵ. This strengthening of Theorem 1.2 pays off in improved estimates for $\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W)$ in the first approach mentioned above (we do not assume that $f_\rho \in W$). The following result is essentially due to [CS] (see [DKPT]). Let W and ρ satisfy (1.1) and (1.5) then for $A \ge A_0(M,D,r)$

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

In Section 2 we prove, for instance, for $r > 1/2$ that for W, ρ satisfying (1.1) and (1.5) we have for $A \ge A_0(M,D,r)$

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

We also prove in Section 2 that one cannot improve the error estimate of order $m^{-1/2}$ in the setting with no assumptions on f_ρ.

It turns out that if we assume that $f_\rho \in W$ we obtain significantly better estimates. We prove the following estimate in Section 3.

Theorem 1.3. *Let $f_\rho \in W$ and let ρ and W satisfy (1.1) and (1.5). Then there exists an estimator f_z such that for $A \ge A_0(M,D,r)$*

$$\text{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \le Am^{-\frac{2r}{1+2r}}\} \ge 1-\exp(-c(M)Am^{\frac{1}{1+2r}}).$$

Let us compare the above estimate with the known result essentially due to [CS] (see [DKPT]).

Theorem 1.4 [CS], [DKPT]. *Assume that ρ and W satisfy (1.1) and (1.5). Suppose that $f_\rho \in W$. Then for $A \ge A_0(M,D,r)$*

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

We note that we have achieved an improved rate of error decay in Theorem 1.3 by replacing $f_{z,W}$ by f_{z,V_ϵ} where V_ϵ is the $\epsilon^{1/2}$-net of W in the $\mathcal{C}$ norm. This raises a natural question of choosing for a given class W an approximation space $\mathcal{H}$ in such a way that the error $\mathcal{E}(f_{z,\mathcal{H}}) - \mathcal{E}(f_\rho)$ is close to optimal error and $\mathcal{H}$ is as simple as possible. In a development of this idea we discuss in Section 3 several interesting settings taken from [DKPT]. In particular, we discuss the case of W satisfying (1.2) instead of (1.1).

Let us make a comment on (1.3). The quantity $\mathcal{E}(f_\rho)$ is an important characteristic of the probability measure ρ. Indeed, for $x \in X$

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

is the variance of the random variable y. Therefore,

$$\mathcal{E}(f_\rho) = \int_X v_\rho(x) d\rho_X$$

is the average (over X) of the variance $v_\rho(x)$. However, the measure ρ is unknown and, therefore, the estimate $\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq \delta$ does not allow us to find an approximate value of $\mathcal{E}(f_\rho)$ because we cannot evaluate $\mathcal{E}(f_z)$. In Section 4 we develop a technique of constructing f_z to approximate $\mathcal{E}(f_\rho)$ by $\mathcal{E}_z(f_z)$. The error estimates for $\mathcal{E}_z(f_z)-\mathcal{E}(f_\rho)$ from Section 4 are weaker in the case $r \geq 1/2$ than the corresponding error estimates for $\mathcal{E}(f_{(z)})-\mathcal{E}(f_\rho)$ obtained in [DKPT] and Section 3 of this paper for other estimator $f_{(z)}$. In the case $r \in (0, 1/2)$ we obtain for $\mathcal{E}_z(f_z) - \mathcal{E}(f_\rho)$ the same order estimates as for $\mathcal{E}(f_{(z)}) - \mathcal{E}(f_\rho)$.

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. 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 $\mathcal{E}(f_z) - \mathcal{E}(f_W)$

In this section we keep notations from Section 1. Denote by $f_\mathcal{H}$ a function from $\mathcal{H}$ that minimizes the error $\mathcal{E}(f)$:

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

For notational simplicity we always assume that such a $f_\mathcal{H}$ exists. Otherwise we would need to take the one that almost minimizes $\mathcal{E}(f)$ over $f \in \mathcal{H}$. The following theorem (see [DKPT]) is essentially contained in [CS].

Theorem 2.1 [CS], [DKPT]. *Assume that W and ρ satisfy (1.1) and (1.5). Then for $A \geq A_0(M, D, r)$*

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

We begin with Theorem 2.2 that is a variant of Theorem 1.2 when we know the behavior of the sequence $E(W) := \{\epsilon_n(W)\}_{n=1}^\infty$. First, we prove an auxiliary result. We will use the $L_\infty := L_\infty(\rho_X)$ norm.

Lemma 2.1. *If $\delta > \eta/(8M)$, $|f_j(x) - y| \le M$ a.e. for $j = 1, 2$, and $\|f_1 - f_2\|_\infty \le \delta$, then*

$$\mathrm{Prob}_{z \in Z^m}\{|L_z(f_1) - L_z(f_2)| \le \eta\} \ge 1 - 2\exp\left(-\frac{m\eta^2}{30M^2\delta^2}\right).$$

Proof. Consider the random variable $\xi = (f_1(x) - y)^2 - (f_2(x) - y)^2$. We use

$$|\xi| \le 2M\|f_1 - f_2\|_\infty \quad \text{a.e.}$$

Therefore, $|\xi - E\xi| \le 4M\delta$ a.e. and the variance V of ξ is at most $4M^2\delta^2$. Applying the Bernstein inequality (1.4) to ξ we get

$$\mathrm{Prob}_{z \in Z^m}\{|L_z(f_1) - L_z(f_2)| \ge \eta\} = \mathrm{Prob}_{z \in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)\right| \ge \eta\right\}$$
$$\le 2\exp\left(-\frac{m\eta^2}{2(4M^2\delta^2 + 4M\delta\eta/3)}\right) \le 2\exp\left(-\frac{m\eta^2}{2(44M^2\delta^2/3)}\right),$$

and Lemma 2.1 follows.

Theorem 2.2. *Assume that ρ, W satisfy (1.5) and W is such that*

$$\sum_{n=1}^\infty n^{-1/2}\epsilon_n(W) < \infty. \tag{2.1}$$

Then for $m\eta^2 \ge 1$ we have

$$\mathrm{Prob}_{z \in Z^m}\{\sup_{f \in W} |L_z(f)| \ge \eta\} \le C(M, E(W))\exp(-c(M)m\eta^2).$$

Proof. It is clear that (2.1) implies that

$$\sum_{j=0}^\infty 2^{j/2}\epsilon_{2^j}(W) < \infty. \tag{2.2}$$

Denote $\delta_j := \epsilon_{2^j}$, $j = 0, 1, \dots$, and consider minimal δ_j-nets $\mathcal{N}_j \subset W$ of W. We will use the notation $N_j := |\mathcal{N}_j|$. Let J be the minimal j satisfying $\delta_j \le \eta/(8M)$. For $j = 1, \dots, J$ we define a mapping A_j that associates with a function $f \in W$ a function $A_j(f) \in \mathcal{N}_j$ closest to f in the $\mathcal{C}$ norm. Then, clearly,

$$\|f - A_j(f)\|_{\mathcal{C}} \le \delta_j.$$

We use the mappings A_j, $j = 1, \dots, J$ to associate with a function $f \in W$ a sequence of functions $f_J, f_{J-1}, \dots, f_1$ in the following way

$$f_J := A_J(f), \quad f_j := A_j(f_{j+1}), \quad j = 1, \dots, J - 1.$$

We introduce an auxiliary sequence

$$\eta_j := (30)^{1/2} M\eta 2^{(j+1)/2}\epsilon_{2^{j-1}}, \quad j = 1, 2, \dots, \tag{2.3}$$

and define $I := I(M, E(W))$ to be the minimal satisfying

$$\sum_{j\geq I} \eta_j \leq \eta/4. \tag{2.4}$$

We now proceed to the estimate of $\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \geq \eta\}$ with m, η satisfying $m\eta^2 \geq 1$. First of all it is not difficult to see ([CS], [DKPT, Proposition 2.2]) that the assumption $\delta_J \leq \eta/(8M)$ implies that if $|L_z(f)| \geq \eta$ then $|L_z(f_J)| \geq \eta/2$. Using this, (2.4), and rewriting

$$L_z(f_J) = L_z(f_J) - L_z(f_{J-1}) + \cdots + L_z(f_{I+1}) - L_z(f_I) + L_z(f_I)$$

we conclude that at least one of the following events occurs:

$$|L_z(f_j) - L_z(f_{j-1})| \geq \eta_j \quad \text{for some} \quad j \in (I, J] \quad \text{or} \quad |L_z(f_I)| \geq \eta/4.$$

Therefore

$$\begin{aligned} \mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \geq \eta\} &\leq \mathrm{Prob}_{z\in Z^m}\{\sup_{f\in \mathcal{N}_I} |L_z(f)| \geq \eta/4\} \\ &+ \sum_{j\in(I,J]}\sum_{f\in\mathcal{N}_j} \mathrm{Prob}_{z\in Z^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\} \\ &\leq \mathrm{Prob}_{z\in Z^m}\{\sup_{f\in \mathcal{N}_I} |L_z(f_I)| \geq \eta/4\} \\ &+ \sum_{j\in(I,J]} N_j \sup_{f\in W} \mathrm{Prob}_{z\in Z^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\}. \end{aligned} \tag{2.5}$$

By our choice of $\delta_j = \epsilon_{2^j}$ we get $N_j \leq 2^{2^j} < e^{2^j}$. Applying Lemma 2.1 we obtain

$$\sup_{f\in W} \mathrm{Prob}_{z\in Z^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\} \leq 2\exp(-\frac{m\eta_j^2}{30M^2\delta_{j-1}^2}).$$

From the definition (2.3) of η_j we get

$$\frac{m\eta_j^2}{30M^2\delta_{j-1}^2} = m\eta^2 2^{j+1}$$

and

$$N_j \exp(-\frac{m\eta_j^2}{30M^2\delta_{j-1}^2}) \leq \exp(-m\eta^2 2^j).$$

Therefore

$$\sum_{j\in(I,J]} N_j \exp(-\frac{m\eta_j^2}{30M^2\delta_{j-1}^2}) \le 2\exp(-m\eta^2 2^I). \tag{2.6}$$

By Theorem 1.2

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in\mathcal{N}_I} |L_z(f)| \ge \eta/4\} \le 2N_I \exp(-\frac{m\eta^2}{C(M)}). \tag{2.7}$$

Combining (2.6) and (2.7) we obtain

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \ge \eta\} \le C(M,E(W))\exp(-c(M)m\eta^2).$$

This completes the proof of Theorem 2.2.

Theorem 2.3. *Assume that ρ, W satisfy (1.5) and W is such that*

$$\sum_{n=1}^{\infty} n^{-1/2}\epsilon_n(W) = \infty.$$

For $\eta > 0$ define $J := J(\eta/M)$ as the minimal j satisfying $\epsilon_{2^j} \le \eta/(8M)$ and

$$S_J := \sum_{j=1}^{J} 2^{(j+1)/2}\epsilon_{2^{j-1}}.$$

Then for m, η satisfying $m(\eta/S_J)^2 \ge 480M^2$ we have

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \ge \eta\} \le C(M,E(W))\exp(-c(M)m(\eta/S_J)^2).$$

Proof. This proof differs from the above proof of Theorem 2.2 only in the choice of an auxiliary sequence $\{\eta_j\}$. Thus we keep notations from the proof of Theorem 2.2. Now, instead of (2.3) we define $\{\eta_j\}$ as follows

$$\eta_j := \frac{\eta}{4}\frac{2^{(j+1)/2}\epsilon_{2^{j-1}}}{S_J}.$$

Proceeding as in the proof of Theorem 2.2 with $I = 1$ we need to check that

$$2^j - \frac{m\eta_j^2}{30M^2\delta_{j-1}^2} \le -2^j\frac{m(\eta/S_J)^2}{480M^2}.$$

Indeed, using the assumption $m(\eta/S_J)^2 \geq 480M^2$ we obtain

$$\frac{m\eta_j^2}{30M^2\delta_{j-1}^2} - 2^j = \frac{m(\eta/S_J)^2}{480M^2}2^{j+1} - 2^j \geq \frac{m(\eta/S_J)^2}{480M^2}2^j.$$

We complete the proof in the same way as in Theorem 2.2.

Remark 2.1. Let $a = \{a_n\}$ be a majorant sequence for $\{\epsilon_n(W)\}$: $\epsilon_n(W) \leq a_n$, $n = 1, 2, \dots$. It is clear that Theorem 2.3 holds with J replaced by $J(a)$ - the minimal j satisfying $a_{2^j} \leq \eta/(8M)$ and with S_J replaced by

$$S_{J(a)} := \sum_{j=1}^{J(a)} 2^{(j+1)/2} a_{2^{j-1}}.$$

We formulate three corollaries of Theorem 2.3. All the proofs are similar. We only proof Corollary 2.3 here.

Corollary 2.1. *Assume ρ, W satisfy (1.5) and $\epsilon_n(W) \leq Dn^{-1/2}$. Then for m, η satisfying $m(\eta/\log(M/\eta))^2 \geq C_1(M,D)$ we have*

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \geq \eta\} \leq C(M,D)\exp(-c(M,D)m(\eta/\log(M/\eta))^2).$$

Corollary 2.2. *Assume ρ, W satisfy (1.5) and $\epsilon_n(W) \leq Dn^{-r}$, $r \in (0,1/2)$. Then for m, η satisfying $m\eta^{1/r} \geq C_1(M,D,r)$ we have*

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in W} |L_z(f)| \geq \eta\} \leq C(M,D,r)\exp(-c(M,D,r)m\eta^{1/r}).$$

Denote by $\mathcal{N}_\delta(W)$ the δ-net of W in the $\mathcal{C}$ norm.

Corollary 2.3. *Assume ρ, W satisfy (1.5) and $\epsilon_n(W) \leq Dn^{-r}$, $r \in (0,1/2)$. Then for m, η, $\delta \geq \eta/(8M)$ satisfying $m\eta^2\delta^{1/r-2} \geq C_1(M,D,r)$ we have*

$$\mathrm{Prob}_{z\in Z^m}\{\sup_{f\in\mathcal{N}_\delta(W)} |L_z(f)| \geq 2\eta\} \leq C(M,D,r)\exp(-c(M,D,r)m\eta^2\delta^{1/r-2}).$$

Proof. We apply Theorem 2.3 to $\mathcal{N}_\delta(W)$. First of all we note that for n such that $\epsilon_n(W) \leq \delta$ we have $\epsilon_n(\mathcal{N}_\delta(W)) = 0$. Also, for n such that $\epsilon_n(W) > \delta$ we have

$$\epsilon_n(\mathcal{N}_\delta(W)) \leq \epsilon_n(W) + \delta \leq 2\epsilon_n(W).$$

We now estimate the S_J from Theorem 2.3. Denote J_δ the minimal j satisfying $\epsilon_{2^j}(W) \leq \delta$ and keep the notation J for the minimal j satisfying $\epsilon_{2^j}(W) \leq \eta/(8M)$. Then it is clear from our assumption $\delta \geq \eta/(8M)$ that $J_\delta \leq J$ and $\epsilon_{2^{j-1}}(\mathcal{N}_\delta(W)) = 0$ for $j > J_\delta$. Therefore,

$$S_J \leq 2\sum_{j=1}^{J_\delta} 2^{(j+1)/2}\epsilon_{2^{j-1}}(W) \leq 2^{3/2+r}D\sum_{j=1}^{J_\delta} 2^{j(1/2-r)} \leq C_1(r)D2^{J_\delta(1/2-r)}.$$

Next,

$$D2^{-r(J_\delta-1)} \ge \epsilon_{2^{J_\delta-1}} > \delta \quad \text{implies} \quad 2^{J_\delta} \le 2(D/\delta)^{1/r}.$$

Thus

$$S_J \le C_1(D,r)(1/\delta)^{\frac{1}{2r}-1}.$$

It remains to apply Theorem 2.3.

Theorem 2.4. *Assume that ρ and W satisfy (1.1) and (1.5). Then we have the following estimates*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W})-\mathcal{E}(f_W)\le\eta\}\ge 1-C(M,D,r)\exp(-c(M)m\eta^2), \tag{2.8}$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W})-\mathcal{E}(f_W)|\le 2\eta\}\ge 1-C(M,D,r)\exp(-c(M)m\eta^2),$$

provided $r>1/2$, $m\eta^2\ge 1$.

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W})-\mathcal{E}(f_W)\le\eta\}\ge 1-C_1(M,D)\exp(-c(M,D)m(\eta/\log(M/\eta))^2),$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W})-\mathcal{E}(f_W)|\le 2\eta\}\ge 1-C_1(M,D)\exp(-c(M,D)m(\eta/\log(M/\eta))^2),$$

provided $r=1/2$, $m(\eta/\log(M/\eta))^2\ge C_2(M,D)$.

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W})-\mathcal{E}(f_W)\le\eta\}\ge 1-C_1(M,D,r)\exp(-c(M,D,r)m\eta^{1/r}),$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W})-\mathcal{E}(f_W)|\le 2\eta\}\ge 1-C_1(M,D,r)\exp(-c(M,D,r)m\eta^{1/r}),$$

provided $r\in(0,1/2)$, $m\eta^{1/r}\ge C_2(M,D,r)$.

Proof. This theorem follows from Theorem 2.2, Corollaries 2.1, 2.2, and the chain of inequalities

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

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

We note that we can take in (2.8) η as small as $\eta=Am^{-1/2}$ and $m^{-1/2}$ is the best rate we can achieve using (2.8). We now prove that in general we cannot estimate $\mathcal{E}(f_{z,W})-\mathcal{E}(f_W)$ with better rate than $m^{-1/2}$.

Proposition 2.1. *There exists a constant $c>0$, a class W consisting of two functions 1 and -1 such that for every $m=2,3,\dots$ there are two measures ρ_0 and ρ_1 such that for any estimator $f_z\in W$ for one of ρ_0, ρ_1 we have*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z)-\mathcal{E}(f_W)\ge m^{-1/2}\}\ge c.$$

Proof. Let $X = [0,1]$, $Y = [-1,1]$. For a given $m \in \mathbb{N}$ we define ρ_0, ρ_1 as follows. For both ρ_0, ρ_1 the ρ_X is the Lebesgue measure on $[0,1]$ (the proof below works for any ρ_X) and for $x \in [0,1]$ we define

$$\rho_0(1|x) = \rho_1(-1|x) = p; \quad \rho_0(-1|x) = \rho_1(1|x) = 1-p$$

with $p = (1+m^{-1/2})/2$. Then

$$f_{\rho_0} = m^{-1/2}; \quad f_{\rho_1} = -m^{-1/2}$$

and

$$(f_{\rho_0})_W = 1; \quad (f_{\rho_1})_W = -1.$$

Let $z = ((x_1,y_1),\dots,(x_m,y_m)) =: (x,y)$, $x = (x_1,\dots,x_m)$, $y = (y_1,\dots,y_m)$. For a fixed $x \in X^m$ we will prove the lower estimate for the probability in Y^m. For a subset $e \subset \{1,\dots,m\}$ we denote by χ_e the vector $y = (y_1,\dots,y_m)$ such that $y_j = 1$ for $j \in e$ and $y_j = -1$ otherwise. For a given estimator f_z consider the following two sets

$$E_1 := \{e \subset \{1,\dots,m\} \quad : \quad f_z = 1 \quad \text{if} \quad z = (x,\chi_e)\},$$

$$E_{-1} := \{e \subset \{1,\dots,m\} \quad : \quad f_z = -1 \quad \text{if} \quad z = (x,\chi_e)\}.$$

Then for the measure ρ_0 we have

$$\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_0})_W) = 0 \quad \text{for} \quad z = (x,\chi_e), \quad e \in E_1$$

$$\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_0})_W) = 4m^{-1/2} \quad \text{for} \quad z = (x,\chi_e), \quad e \in E_{-1}.$$

Similarly for the measure ρ_1 we have

$$\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_1})_W) = 4m^{-1/2} \quad \text{for} \quad z = (x,\chi_e), \quad e \in E_1$$

$$\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_1})_W) = 0 \quad \text{for} \quad z = (x,\chi_e), \quad e \in E_{-1}.$$

The probability of realization of $y = \chi_e$ in the case of measure ρ_0 is equal to $p^{|e|}(1-p)^{m-|e|}$ and in the case of measure ρ_1 is equal to $p^{m-|e|}(1-p)^{|e|}$. Therefore in the case of ρ_0 we have

$$\mathrm{Prob}_{y\in Y^m}\{\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_0})_W) = 4m^{-1/2}\} = \sum_{e\in E_{-1}} p^{|e|}(1-p)^{m-|e|}$$

and in the case of ρ_1

$$\mathrm{Prob}_{y\in Y^m}\{\mathcal{E}(f_z) - \mathcal{E}((f_{\rho_1})_W) = 4m^{-1/2}\} = \sum_{e\in E_1} p^{m-|e|}(1-p)^{|e|}$$

We will prove that for $p = (1+m^{-1/2})/2$ we have

$$\Sigma := \sum_{e\in E_{-1}} p^{|e|}(1-p)^{m-|e|} + \sum_{e\in E_1} p^{m-|e|}(1-p)^{|e|} \geq c_1 > 0 \tag{2.9}$$

with absolute constant c_1. This implies Proposition 2.1. We restrict summation in both sums from (2.9) to those e with $m/2 - m^{1/2} \le |e| \le m/2 + m^{1/2}$. For such an e we have

$$p^{|e|}(1-p)^{m-|e|} = 2^{-m}(1+m^{-1/2})^{|e|}(1-m^{-1/2})^{m-|e|}$$

$$\ge 2^{-m}(1-m^{-1})^{m/2}(1-m^{-1/2})^{2m^{1/2}} \ge c_2 2^{-m}.$$

Therefore,

$$\Sigma \ge c_2 2^{-m} \sum_{|m/2-k|\le m^{1/2}} C_m^k \ge c_1 > 0.$$

We note that Proposition 2.1 is based on a probabilistic argument for $\rho(y|x)$ and reflects the fact that saturation of the error estimate at the level $m^{-1/2}$ is due to the probabilistic feature of the problem. We will show in the following proposition that the corresponding lower estimate in the case $r \in (0,1/2]$ can be obtained for the Dirac measure $\rho(y|x)$. Thus in this case ($r \in (0,1/2]$) the lower estimate is entailed by the deterministic (in a certain sense) feature of the problem.

Proposition 2.2. *For any $r \in [0,1/2]$ and for every $m \in \mathbb{N}$ there is $W \subset U(L_\infty([0,1])$ satisfying $\epsilon_n(W, L_\infty) \le n^{-r}$ for $n \in \mathbb{N}$ such that for every estimator $f_z \in W$ there is a ρ such that*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W) \ge m^{-r}/4\} \ge 1/7.$$

Proof. Let as above $X = [0,1]$, $Y = [-1,1]$, ρ_X is the Lebesgue measure on $[0,1]$. Define

$$\Gamma = \{\gamma = (\gamma_1, \dots, \gamma_{2m}) : \gamma_i \in \{1,-1\}\,(i = 1,\dots,2m)\}.$$

For $\gamma \in \Gamma$, $x \in [0,1)$ let

$$g_\gamma(x) = \gamma_{[2mx+1]},$$

$$f_\gamma(x) = g_\gamma(x)[2mx+1]^{-r}/2,$$

where $[u]$ is the greatest integer not exceeding u. Let $W = \{f_\gamma : \gamma \in \Gamma\}$. For any $n \in \mathbb{N}$, $n \le 2m$, we can find $\Gamma_n \subset \Gamma$ so that $\#\Gamma_n = 2^n$ and

$$\{(\gamma_1, \dots, \gamma_n) : \gamma \in \Gamma_n\} = \{1,-1\}^n.$$

This means that for every $\gamma \in \Gamma$ there is $\tilde\gamma \in \Gamma_n$ such that $\gamma_i = \tilde\gamma_i$ for $i = 1,\dots,n$ and thus,

$$\|f_\gamma - f_{\tilde\gamma}\|_\infty \le (n+1)^{-r}.$$

Therefore, $\epsilon_n(W, L_\infty) \le n^{-r}$ as required.

We will consider a set of probability measures ρ. It will be convenient for us to use a probabilistic interpretation of this set of measures. Assume that f_ρ is equal to one of the functions g_γ, $\gamma \in \Gamma$, with equal probability. So, for each $\gamma \in \Gamma$

$$\mathrm{Prob}\{f_\rho = g_\gamma\} = 2^{-2m}.$$

We define $\rho(y|x)$ as the Dirac measure: $y_i = f_\rho(x_i)$. Clearly, if $f_\rho = g_\gamma$ then $(f_\rho)_W = f_\gamma$. For every estimator $f_z = f_{\gamma(z)}$ we have for the i such that $\gamma_i(z) \neq \gamma_i$,

$$\begin{aligned}&(f_z(x) - f_\rho(x))^2 - ((f_\rho)_W(x) - f_\rho(x))^2\\ &= \left(1 + i^{-r}/2\right)^2 - \left(1 - i^{-r}/2\right)^2 = 2i^{-r},\end{aligned}$$

where $[2mx + 1] = i$. Therefore,

$$\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W) = \sum_{i:\gamma_i(z)\neq\gamma_i} \frac{1}{m} i^{-r}. \tag{2.10}$$

It is easy to conclude from here that always

$$|\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W)| \leq \sum_{i=1}^{2m} \frac{1}{m} i^{-r}. \tag{2.11}$$

Denote

$$I(z) := \{[2mx_j + 1] : j = 1, \dots, m\}.$$

If $[2mx_j + 1] = i$, then $\gamma_i = y_j$. Thus, for $i \in I(z)$ the value γ_i is determined, and it is natural to consider $\gamma_i(z) = \gamma_i$. However, for $i \notin I(z)$ the probability that $\gamma_i(z) \neq \gamma_i$ is $1/2$. Hence, by (2.10),

$$\mathbb{E}\left(\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W)|z\right) = \frac{1}{2} \sum_{i\notin I(z)} \frac{1}{m} i^{-r}.$$

Next, for every $i \in \{1, \dots, 2m\}$ the probability of the event $i \notin I(z)$ is equal to $(1 - 1/(2m))^m \geq 1/2$. Therefore,

$$\mathbb{E}\left(\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W)\right) \geq \frac{1}{4} \sum_{i=1}^{2m} \frac{1}{m} i^{-r}. \tag{2.12}$$

By (2.11) and (2.12) we get

$$\text{Prob}\{\mathcal{E}(f_z) - \mathcal{E}((f_\rho)_W) \geq \frac{1}{8} \sum_{i=1}^{2m} \frac{1}{m} i^{-r}\} \geq \frac{1}{7}.$$

Taking into account that

$$\sum_{i=1}^{2m} \frac{1}{m} i^{-r} > \frac{1}{m} \int_1^{2m+1} u^{-r} du > \frac{(2m)^{1-r}}{m(1-r)} > 2m^{-r}$$

we complete the proof of the proposition.

3. Estimating $\mathcal{E}(f_z) - \mathcal{E}(f_\rho)$

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

Theorem 3.1 [CS]. *Suppose that either W is a compact and convex subset of $\mathcal{C}(X)$ or W is a compact subset of $\mathcal{C}(X)$ and $f_\rho \in W$. Assume that ρ, W satisfy (1.5). Then, for all $\epsilon > 0$*

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

We will need the following theorem from [DKPT] in a style of Theorem 3.1.

Theorem 3.2 [DKPT]. *Let W be a compact subset of $\mathcal{C}(X)$. Assume that ρ, W satisfy (1.5). Then, for all $\epsilon > 0$*

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

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

The following theorem from [DKPT] is essentially contained in [CS].

Theorem 3.3 [CS], [DKPT]. *Assume that ρ, W satisfy (1.1) and (1.5). Suppose that either W is convex or $f_\rho \in W$. Then for $A \ge A_0(M,D,r)$*

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

We prove here in the case $f_\rho \in W$ a stronger estimate than in the above theorem.

Theorem 3.4. *Let $f_\rho \in W$ and let ρ, W satisfy (1.1) and (1.5). Then there exists an estimator f_z such that for $A \ge A_0(M,D,r)$*

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

Proof. We have assumed that W is such that

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

Then

$$N(W,\epsilon) \le 2^{(D/\epsilon)^{1/r}+1}. \tag{3.1}$$

We choose $\epsilon = Am^{-\frac{2r}{1+2r}}$ and define V_ϵ to be a $\epsilon^{1/2}$-net of W in the $\mathcal{C}$ norm. Then by (3.1)

$$|V_\epsilon| \le 2^{(D^2/\epsilon)^{1/(2r)}+1}.$$

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

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

We now estimate $\mathcal{E}(f_{z,V_\epsilon}) - \mathcal{E}(f_\rho)$. Let $f^* \in V_\epsilon$ be such that

$$\|f_\rho - f^*\|_{\mathcal{C}} \le \epsilon^{1/2} \le A^{1/2} m^{-\frac{r}{1+2r}}.$$

Then

$$\mathcal{E}(f^*) - \mathcal{E}(f_\rho) = \int_X (f^*(x) - f_\rho(x))^2 d\rho_X \le Am^{-\frac{2r}{1+2r}}. \tag{3.2}$$

In particular, this implies that $\mathcal{E}(f_{V_\epsilon}) - \mathcal{E}(f_\rho) \le Am^{-\frac{2r}{1+2r}} = \epsilon$. We have

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

Next,

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

Taking into account the choice of $\epsilon = Am^{-\frac{2r}{1+2r}}$ and $\mathcal{E}(f_{V_\epsilon}) - \mathcal{E}(f_\rho) \le \epsilon$ we get from Theorem 3.2 for $A > A_0(M,D,r)$

$$\begin{aligned} \mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,V_\epsilon}) - \mathcal{E}(f^*) \le Am^{-\frac{2r}{1+2r}}\} \\ \ge 1 - \exp(-c(M)Am^{\frac{1}{1+2r}}). \end{aligned} \tag{3.3}$$

Using (3.2) we obtain from here

$$\begin{aligned} \mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,V_\epsilon}) - \mathcal{E}(f_\rho) \le 2Am^{-\frac{2r}{1+2r}}\} \\ \ge 1 - \exp(-c(M)Am^{\frac{1}{1+2r}}). \end{aligned} \tag{3.4}$$

This completes the proof of Theorem 3.4.

Remark 3.1. The above proof of Theorem 3.4 works also in a little more general situation. Assume instead of (1.1) that W satisfies

$$\epsilon_n(W) \le D((\ln n)^b/n)^r, \quad n = 2,3,\dots, \quad W \subset DU(\mathcal{C}). \tag{3.5}$$

Then similarly to Theorem 3.4 we obtain that there exists an estimator f_z such that for $A \ge A_0(M,D,r,b)$

$$\begin{aligned} \mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \le 2A((\ln m)^b/m)^{\frac{2r}{1+2r}}\} \\ \ge 1 - \exp(-c(M)Am^{\frac{1}{1+2r}}(\ln m)^{\frac{2br}{1+2r}}), \quad m = 2,3,\dots. \end{aligned} \tag{3.6}$$

We compare Theorem 3.4 with two theorems from [DKPT]. Let us assume that W satisfies the following estimates for the Kolmogorov widths (see (1.2))

$$d_n(W,\mathcal{C}) \le Kn^{-r}, \quad n = 1,2,\dots; \quad W \subset KU(\mathcal{C}). \tag{3.7}$$

Then by Carl's [C] inequality

$$\epsilon_n(W) \le Dn^{-r}, \quad n = 1,2,\dots. \tag{3.8}$$

Therefore, for this class we have the estimate as above in Theorem 3.4. Thus Theorem 3.4 gives a slightly better estimate that in the following theorem from [DKPT]. However, we point out that the constructions of estimators that provide the corresponding error estimates are different. It seems that the construction from [DKPT] is simpler.

Theorem 3.5 [DKPT]. *Let $f_\rho \in W$ and let ρ, W satisfy (1.2) and (1.5). Then there exists an estimator f_z such that for $A \geq A_0(M,K,r)$*

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

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]):

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

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

$$d_n(W,\mathcal{C},n^{an})\leq Kn^{-r},\quad n=1,2,\dots;\quad W\subset KU(\mathcal{C}). \tag{3.9}$$

Then by [T5]

$$\epsilon_n(W)_{\mathcal{C}}\leq C(r)K(\ln n/n)^r,\quad n=2,3,\dots.$$

For this class we have the estimate (3.6). It is clear that a class satisfying (3.9) is wider than the class satisfying (1.2). The following result has been obtained in [DKPT].

Theorem 3.6 [DKPT]. *Let $f_\rho \in W$ and let ρ, W satisfy (1.5) and (3.9). Then there exists an estimator f_z such that for $A \geq A_0(M,K,r)$*

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

4. Estimating $\mathcal{E}_z(f_z)-\mathcal{E}(f_\rho)$

The following theorem is essentially contained in [CS].

Theorem 4.1 [CS]. *Assume that ρ, W satisfy (1.1) and (1.5). Then for $A\geq A_0(M,D,r)$*

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

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

First, we discuss estimates that follow directly from the results of Section 2. Theorem 2.4 implies the following estimates.

Theorem 4.2. *Let ρ, W satisfy (1.1) and (1.5). Then for $A \geq A_0(M,D,r)$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \leq A(\ln m/m)^{1/2}\} \geq 1 - C(M,D,r)m^{-c(M)A},$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W}) - \mathcal{E}(f_W)| \leq 2A(\ln m/m)^{1/2}\} \geq 1 - C(M,D,r)m^{-c(M)A},$$

provided $r > 1/2$,

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \leq A((\ln m)^3/m)^{1/2}\} \geq 1 - C(M,D)m^{-c(M,D)A},$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W}) - \mathcal{E}(f_W)| \leq 2A((\ln m)^3/m)^{1/2}\} \geq 1 - C(M,D)m^{-c(M,D)A},$$

provided $r = 1/2$,

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_{z,W}) - \mathcal{E}(f_W) \leq A(\ln m/m)^r\} \geq 1 - C(M,D,r)m^{-c(M,D,r)A},$$

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_{z,W}) - \mathcal{E}(f_W)| \leq 2A(\ln m/m)^r\} \geq 1 - C(M,D,r)m^{-c(M,D,r)A},$$

provided $r \in (0,1/2]$.

Second, we discuss estimates that can be obtained by combining results from Section 3 with results from Section 2.

Theorem 4.3. *Let $f_\rho \in W$ and let ρ, W satisfy (1.1) and (1.5). Then there exists an estimator f_z such that for $A \geq A_0(M,D,r) \geq 2$*

$$\mathrm{Prob}_{z\in Z^m}\{\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 2Am^{-\frac{2r}{1+2r}}\} \geq 1 - \exp(-c(M)Am^{\frac{1}{1+2r}}). \tag{4.1}$$

Also

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_z) - \mathcal{E}(f_\rho)| \leq 3A(\ln m/m)^{1/2}\} \geq 1 - C(M,D,r)m^{-c(M)A^2}, \tag{4.2}$$

provided $r > 1/2$,

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_z) - \mathcal{E}(f_\rho)| \leq 3A((\ln m)^3/m)^{1/2}\} \geq 1 - C(M,D)m^{-c(M,D)(A/\log A)^2}, \tag{4.3}$$

provided $r = 1/2$,

$$\mathrm{Prob}_{z\in Z^m}\{|\mathcal{E}_z(f_z) - \mathcal{E}(f_\rho)| \leq 4A(\ln m)^{1/2}m^{-\frac{2r}{1+2r}}\} \geq 1 - C(M,D,r)m^{-c(M,D,r)A^{1+\frac{1}{2r}}}, \tag{4.4}$$

for $m \geq C(A,M)$ provided $r \in (0,1/2)$.

Proof. Theorem 3.4 provides an estimator $f_z \in W$ that satisfies (4.1). We will prove (4.2) and (4.3) for any estimator $f_z \in W$ satisfying (4.1). First, we consider the case $r > 1/2$. We use Theorem 2.2 with $\eta = A(\ln m/m)^{1/2}$ applied to the class W. Then we obtain

$$|\mathcal{E}_z(f_z) - \mathcal{E}(f_z)| \leq A(\ln m/m)^{1/2} \tag{4.5}$$

with probability at least $1 - C(M,D,r)m^{-c(M)A^2}$. Combining (4.1) and (4.5) we get (4.2).

Second, we consider the case $r = 1/2$. We use Corollary 2.1 with $\eta = A((\ln m)^3/m)^{1/2}$ applied to the class W. Then we obtain

$$|\mathcal{E}_z(f_z) - \mathcal{E}(f_z)| \le A((\ln m)^3/m)^{1/2} \tag{4.6}$$

with probability at least $1 - C(M,D,r)m^{-c(M,D)(A/\log A)^2}$ for $A \ge 2$. Combining (4.1) and (4.6) we get (4.3).

We proceed to the case $r \in (0,1/2)$. Contrary to the above two cases we will use a specific form of the estimator f_z from Theorem 3.4. The estimator f_z from Theorem 3.4 is

$$f_z := f_{z,V_\epsilon} = \arg\min_{f\in V_\epsilon} \mathcal{E}_z(f)$$

where V_ϵ is a $\epsilon^{1/2}$-net of W in the $\mathcal{C}$ norm, $\epsilon = Am^{-\frac{2r}{1+2r}}$. We now estimate $L_z(f_z)$. We note that $V_\epsilon \subset W$ and therefore for all ϵ the set V_ϵ satisfies (1.1) (with D replaced by $2D$) if W satisfies (1.1). We use Corollary 2.3 with $\eta = A(\ln m)^{1/2}m^{-\frac{2r}{1+2r}}$, $\delta = \epsilon^{1/2} = A^{1/2}m^{-\frac{r}{1+2r}}$ applied to the compact V_ϵ. Then for $m \ge C(M,A)$ we have $\delta \ge \eta/(8M)$ and $m\eta^2\delta^{\frac{1}{r}-2} \ge C_1(M,D,r)$. Using Corollary 2.3 we obtain

$$|\mathcal{E}_z(f_z) - \mathcal{E}(f_z)| \le 2A(\ln m)^{1/2}m^{-\frac{2r}{1+2r}} \tag{4.7}$$

with probability at least $1 - C(M,D,r)m^{-c(M,D,r)A^{1+\frac{1}{2r}}}$. Combining (4.1) and (4.7) we get (4.4).

References

[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.

[DKPT] R. DeVore, G. Kerkyacharian, D. Picard, V. Temlyakov, *On Mathematical Methods of Learning*, Manuscript (2003), 1–23.

[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.