APPROXIMATION AND PROBABILITY
BANACH CENTER PUBLICATIONS, VOLUME 72
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSZAWA 2006

OPTIMAL ESTIMATORS IN LEARNING THEORY

V. N. TEMLYAKOV

Department of Mathematics, University of South Carolina
Columbia, SC 29208, U.S.A.
E-mail: temlyak@math.sc.edu

This paper is dedicated to the 70th birthday of Zbigniew Ciesielski

Abstract. This paper is a survey of recent results on some problems of supervised learning in the setting formulated by Cucker and Smale. 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 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 well the regression function f_ρ of an unknown Borel probability measure ρ defined on $Z = X \times Y$. We assume that (x_i, y_i), $i = 1, \dots, m$, are indepent and distributed according to ρ. We discuss a problem of finding optimal (in the sense of order) estimators for different classes Θ (we assume $f_\rho \in \Theta$). It is known from the previous works that the behavior of the entropy numbers $\epsilon_n(\Theta, B)$ of Θ in a Banach space B plays an important role in the above problem. The standard way of measuring the error between a target function f_ρ and an estimator $f_{\mathbf{z}}$ is to use the $L_2(\rho_X)$ norm (ρ_X is the marginal probability measure on X generated by ρ). The usual way in regression theory to evaluate the performance of the estimator $f_{\mathbf{z}}$ is by studying its convergence in expectation, i.e. the rate of decay of the quantity $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$ as the sample size m increases. Here the expectation is taken with respect to the product measure ρ^m defined on Z^m. A more accurate and more delicate way of evaluating the performance of $f_{\mathbf{z}}$ has been pushed forward in [CS]. In [CS] the authors study the probability distribution function

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \ge \eta\}$$

instead of the expectation $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$. In this survey we mainly discuss the optimization problem formulated in terms of the probability distribution function.

2000 *Mathematics Subject Classification*: 62G05, 62G08.
This research was supported by the National Science Foundation Grant DMS 0200187.
The paper is in final form and no version of it will be published elsewhere.

1. Introduction. Notations. Settings. This paper is a survey of recent results on supervised learning. 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 $y \in Y$. This is a big area of research both in nonparametric statistics and in learning theory. In this paper we confine ourselves to recent results obtained in a direction of further development of the settings and results from the fundamental paper of Cucker and Smale [CS]. In this paper we illustrate how methods of approximation theory can be used in learning theory. We begin our discussion with a very brief survey of different settings that are close to the setting of our main interest.

1. *Approximation theory. Recovery of functions.* Deterministic model: given

$$\mathbf{z} := ((x_1, y_1), \dots, (x_m, y_m)) \quad : \quad y_i = f(x_i), \quad i = 1, \dots, m, \quad f \in \Theta.$$

Recover $f \in \Theta$ (find an approximant of f). Error of approximation is measured in some norm $\|\cdot\|$. Usually it is the L_p norm, $1 \le p \le \infty$, with respect to the Lebesgue measure on a given domain X.

2. *Statistics. Regression theory.*

a) Fixed design model: given

$$\mathbf{z} := ((x_1, y_1), \dots, (x_m, y_m)) \quad : \quad y_i = f(x_i) + \epsilon_i, \quad x_1, \dots, x_m \text{ fixed},$$

ϵ_i independent identically distributed (i.i.d.), $E\epsilon_i = 0$, $f \in \Theta$.

Find an approximant for f (estimator $\hat{f}$). The unknown function f is called the regression function. Error is measured by expectation $E(\|f - \hat{f}\|^2)$ of some of the standard norms.

b) Random design model: given

$$\mathbf{z} := ((x_1, y_1), \dots, (x_m, y_m)) \quad : \quad y_i = f(x_i) + \epsilon_i,$$

$x_1, \dots, x_m$ random, i.i.d.; ϵ_i i.i.d. (independent of x_i), $E\epsilon_i = 0$, $f \in \Theta$. Find an estimator $\hat{f}$ for f. Error is measured by expectation $E(\|f - \hat{f}\|^2)$.

c) Distribution-free theory of regression.

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 $\rho(y|x)$, the 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 minimizes the error $\mathcal{E}(f)$. It is known in statistics as the *regression function* of ρ. Given: (x_i, y_i), $i = 1, \dots, m$, independent identically distributed according to ρ, $|y| \le M$ a.e. Find an estimator $\hat{f}$ for f_ρ. Error: $E(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)})$. Assume $f_\rho \in \Theta$.

For a class Θ consider

$$E(\Theta, m, \hat{f}) := \sup_{f_\rho \in \Theta} E(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)}), \qquad E(\Theta, m) := \inf_{\hat{f}} E(\Theta, m, \hat{f}).$$

3. *Learning theory.* This is a vast area of research with a wide range of different settings. In this paper we only discuss a development of a setting from [CS]. For results in other settings we recommend a fundamental book of V. Vapnik [V] and a nice survey on the classification problem by G. Lugosi [L]. Our setting is similar to the setting of the distribution-free regression problem. The 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 (or its projection) well with high probability. We assume that (x_i, y_i), $i = 1, \dots, m$ are independent and distributed according to ρ. Similarly to the distribution-free theory of regression we measure the error in the $L_2(\rho_X)$ norm. This differs the distribution-free theory of regression and our setting of learning theory from classical nonparametric statistics. One can find a discussion of relations between the fixed design model, the random design model, and the distribution-free theory of regression in the recent book [GKKW] (see also [VG], [BM1]). Here we only mention that the problem of learning theory that we discuss in this paper can be rewritten in the form

$$y_i = f_\rho(x_i) + \epsilon_i, \quad \epsilon := y - f_\rho(x),$$

close to the form of the random design model. However, in our setting we are not assuming that ϵ and x are independent. While the theories of fixed and random design models do not directly apply to our setting, they utilize several of the same techniques we shall encounter such as the use of entropy and the construction of estimators through minimal risk.

We note that a standard setting in the distribution-free theory of regression (see [GKKW]) involves the expectation as a measure of quality of an estimator. An important new feature of the setting in learning theory formulated in [CS] is the following. They propose to study systematically the probability distribution function

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \ge \eta\}$$

instead of the expectation. There are several important ingredients in mathematical formulation of the learning problem. In our formulation we follow the way that has become standard in approximation theory and based on the concept of *optimal method*.

We begin with a class $\mathcal{M}$ of admissible measures ρ. Usually, we impose restrictions on ρ in the form of restrictions on the regression function f_ρ: $f_\rho \in \Theta$. Then the first step is to find an optimal estimator for a given class Θ of priors (we assume $f_\rho \in \Theta$). In regression theory the usual way to evaluate the performance of the estimator $f_{\mathbf{z}}$ is by studying its convergence in expectation, i.e. the rate of decay of the quantity $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$ as the sample size m increases. Here the expectation is taken with respect to the product measure ρ^m defined on Z^m. We note that $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)}$. As we already mentioned above a more accurate and more delicate way of evaluating the performance of $f_{\mathbf{z}}$ has been pushed forward in [CS]. In this paper we concentrate on a discussion of results on the probability distribution function.

An important question in finding an optimal $f_{\mathbf{z}}$ is the following. How to describe the class Θ of priors? In other words, what characteristics of Θ govern, say, the optimal rate

of decay of $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$ for $f_\rho \in \Theta$? Previous and recent works in statistics and learning theory (see [B], [BM2], [BM3], [CS], [DKPT1], [DKPT2], [GKKW], [KT1], [KT2], [L], [V], [VG]) indicate that the compactness characteristics of Θ play a fundamental role in the above problem. It is convenient for us to express compactness of Θ in terms of the entropy numbers. In this survey we discuss the classical concept of entropy and the concept of tight entropy. We note that some other concepts of entropy, for instance, entropy with bracketing, proved to be useful in the theory of empirical processes and nonparametric statistics (see [VG], [BM2], [V]). There is a concept of VC dimension that plays a fundamental role in the problem of pattern recognition and classification [V]. This concept is also useful in describing compactness characteristics of sets. We do not discuss this concept here because we have no new results in this direction.

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

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

where $U(B)$ is the unit ball of Banach space B. We denote $N(\Theta, \epsilon, B)$ the covering number that is the minimal number of balls of radius ϵ needed for covering Θ. The corresponding ϵ-net is denoted by $\mathcal{N}_\epsilon(\Theta, B)$. In the papers [CS], [DKPT1], [DKPT2], [KT1] in the most cases the space $\mathcal{C} := \mathcal{C}(X)$ of continuous functions on a compact $X \subset \mathbb{R}^d$ has been taken as a Banach space B. This allowed us to formulate all results with assumptions on Θ independent of ρ. In [KT2] and [BCDDT] we obtain some results for $B = L_2(\rho_X)$. On the one hand we weaken assumptions on the class Θ and on the other hand this results in the use of ρ_X in the construction of an estimator. Thus, we have a tradeoff between treating wider classes and building estimators that are independent of ρ_X. We note that in practice we often do not know the ρ_X. Thus, it is very desirable to build estimators independent of ρ_X. In statistics this type of regression problem is referred to as *distribution-free*. A recent survey on distribution-free regression theory is provided in the book [GKKW].

In Sections 2 and 3 of this paper we always assume that the unknown measure ρ satisfies the condition $|y| \le M$ (or a little weaker $|y| \le M$ a.e. with respect to ρ_X) with some fixed M. Then it is clear that for f_ρ we have $|f_\rho(x)| \le M$ for all x (for almost all x). Therefore, it is natural to assume that a class Θ of priors where f_ρ belongs is embedded into the $\mathcal{C}(X)$-ball (L_∞-ball) of radius M. We make this assumption in all theorems of Sections 2 and 3 without formulating the assumption.

In [DKPT1], [DKPT2], [KT1] the restrictions on a class Θ have been imposed in the following forms:

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

or

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

Here, $d_n(\Theta, B)$ is the Kolmogorov width. Kolmogorov's n-width for the centrally symmetric compact set Θ in the Banach space B is defined as follows

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

where $\inf_L$ is taken over all n-dimensional linear subspaces of B. In [KT2] we impose a weaker restriction

$$\epsilon_n(\Theta, L_2(\rho_X)) \le Dn^{-r}, \quad n = 1, 2, \dots, \quad \Theta \subset DU(L_2(\rho_X)). \tag{1.3}$$

We have already mentioned above that the study of the probability distribution function $\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \ge \eta\}$ is a more difficult and delicate problem than the study of the expectation $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$. We encounter this difficulty even at the level of formulation of a problem. The reason for this is that the probability distribution function provides control of two characteristics: η, the error of estimation, and $1 - \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \ge \eta\}$, the confidence of the error η. Therefore, we need a mathematical formulation of the above discussed problems of optimal estimators.

We propose (see [DKPT2]) 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 the class of all estimators, the 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 the case $\mathbb{E}(m)$ is the set of all estimators, $m = 1, 2, \dots$, we write $\mathbf{AC}_m(\mathcal{M}, \eta)$.

In Section 2 we discuss results on $\mathbf{AC}_m(\mathcal{M}, \mathbb{E}, \eta)$ with $\mathcal{M} = \mathcal{M}(\Theta) := \{\rho : f_\rho \in \Theta\}$. In this case we write $\mathbf{AC}_m(\mathcal{M}(\Theta), \mathbb{E}, \eta) =: \mathbf{AC}_m(\Theta, \mathbb{E}, \eta)$. Thus Section 2 is devoted to the study of priors on f_ρ in the form $f_\rho \in \Theta$. Sometimes this setting is referred to as *proper function learning problem*.

It is clear from the definition of $E(\Theta, m)$ and $\mathbf{AC}_m(\Theta, \eta)$ that

$$\int_0^\infty \mathbf{AC}_m(\Theta, \eta^{1/2}) d\eta \le E(\Theta, m), \tag{1.4}$$

and for ρ, Θ satisfying $|y| \le M$, $\Theta \subset MU(\mathcal{C}(X))$

$$E(\Theta, m) \le \min_\eta (\eta^2 + 4M^2 \mathbf{AC}_m(\Theta, \eta)). \tag{1.5}$$

One of the important variants of the learning problem formulated in [CS] is the following. We now do not impose any restrictions on ρ, except $|y| \le M$ a.e. and instead of estimating the regression function f_ρ we estimate a projection $(f_\rho)_W$ of f_ρ onto a compact set W of our choice. Sometimes this setting is referred to as *improper function learning problem*. Similarly to the above case ($f_\rho \in \Theta$) we introduce the corresponding accuracy confidence function

$$\mathbf{AC}^p_m(W, \mathbb{E}, \eta) := \inf_{E_m \in \mathbb{E}(m)} \sup_\rho \rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}((f_\rho)_W) \ge \eta^2\}.$$

In the case $\mathbb{E}(m)$, $m = 1, 2, \dots$, is a collection of all estimators $E_m : \mathbf{z} \to f_{\mathbf{z}} \in W$ we drop $\mathbb{E}$ from the notation. We note that in the case of convex W we have for any $f \in W$

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

We discuss related results in Section 3.

In Section 4 we discuss an important statistical problem of how well the *empirical error (risk)* of f

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

can approximate the actual error $\mathcal{E}(f)$. This problem is related to the concept of the Glivenko-Cantelli sample complexity.

Section 5 contains a probabilistic inequality that we use in the discussion in Section 2. This inequality might be of an independent interest.

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 $b_n \ll a_n$.

2. Prior on f_ρ in the form $f_\rho \in \Theta$. We begin with the lower estimate of the accuracy confidence function from [DKPT2]. We shall establish lower bounds in terms of a certain variant of the Kolmogorov entropy of Θ which we shall call *tight entropy*. This type of entropy has been used to prove lower bounds in approximation theory. Also, a similar type of entropy was used by Yang and Barron [YB] in statistical estimation. The entropy measure that we shall use is in general different from the Kolmogorov entropy, but, for classical smoothness sets Θ, it is equivalent to the Kolmogorov entropy and therefore our lower bounds will apply in these classical settings.

For a compact Θ in a Banach space B we define the *packing numbers* as

$$P(\Theta, \delta) := P(\Theta, \delta, B) := \sup\{N : \exists\, f_1, ..., f_N \in \Theta,\ \delta \le \|f_i - f_j\|_B,\ \forall i \ne j\}. \tag{2.1}$$

It is well known [P] and easy to check that $N(\Theta, \delta, B) \le P(\Theta, \delta, B)$. The *tight packing numbers* are defined as follows. Let $1 \le c_1 < \infty$ be a fixed real number. We define the tight packing numbers as

$$\bar{P}(\Theta, \delta) := \bar{P}(\Theta, \delta, c_1, B) := \sup\{N : \exists f_1, \dots, f_N \in \Theta,\ \delta \le \|f_i - f_j\|_B \le c_1\delta,\ \forall i \ne j\}. \tag{2.2}$$

It is clear that $\bar{P}(\Theta, \delta, c_1, B) \le P(\Theta, \delta, B)$.

We let μ be any Borel measure defined on X and let $\mathcal{M}(\Theta, \mu)$ denote the set of all $\rho \in \mathcal{M}(\Theta)$ such that $\rho_X = \mu$, $|y| \le 1$. As above $\mathcal{M}(\Theta) = \{\rho : f_\rho \in \Theta\}$. We specify $B = L_2(\mu)$ and assume that $\Theta \subset L_2(\mu)$. We will use the abbreviated notation $\bar{P}(\delta) := \bar{P}(\Theta, \delta, c_1, L_2(\mu))$.

Let us fix any set Θ and any Borel measure μ defined on X. We set $\mathcal{M} := \mathcal{M}(\Theta, \mu)$ as defined above. We also take $1 < c_1$ in an arbitrary way but then fix this constant. For any fixed $\delta > 0$, we let $\{f_i\}_{i=1}^{\bar{P}}$, with $\bar{P} := \bar{P}(\delta)$, be a net of functions satisfying (2.2). To

each f_i, we shall associate the measure

$$d\rho_i(x,y) := (a_i(x)d\delta_1(y) + b_i(x)d\delta_{-1}(y))d\mu(x),$$

where $a_i(x) := (1+f_i(x))/2$, $b_i(x) := (1-f_i(x))/2$ and $d\delta_\xi$ denotes the Dirac delta with unit mass at ξ. Notice that $(\rho_i)_X = \mu$ and $f_{\rho_i} = f_i$ and hence each ρ_i is in $\mathcal{M}(\Theta,\mu)$.

We have the following theorem.

THEOREM 2.1 ([DKPT2]). *Let $1 < c_1$ be a fixed constant. Suppose that Θ is a subset of $L_2(\mu)$ with tight packing numbers $\bar{P} := \bar{P}(\delta)$. In addition suppose that for $\delta = 2\eta > 0$, the net of functions $\{f_i\}_{i=0}^{\bar{P}}$ in (2.2) satisfies $\|f_i\|_{\mathcal{C}(X)} \le 1/4$, $i = 1,\dots,\bar{P}$. Then for any estimator $f_{\mathbf{z}}$ we have for some $i \in \{1,\dots,\bar{P}\}$*

$$\rho_i^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_i\|_{L_2(\mu)} \ge \eta\} \ge \min(1/2, (\bar{P}(2\eta)-1)^{1/2} e^{-8c_1^2 m\eta^2 - 3/e}),$$
$$\forall \eta > 0,\ m = 1,2,\dots.$$

The proof of Theorem 2.1 is given in [DKPT2]. This proof uses the concept of the Kullback-Leibler information. Given two probability measures dP and dQ defined on the same space and such that dP is absolutely continuous with respect to dQ, we write $dP = gdQ$ and define

$$\mathcal{K}(P,Q) := \int \ln g dP = \int g \ln g dQ.$$

If dP is not absolutely continuous with respect to dQ then $\mathcal{K}(P,Q) := \infty$.

It is obvious that

$$\mathcal{K}(P^m, Q^m) = m\mathcal{K}(P,Q).$$

The use of Kullback-Leibler information is well known in statistics and goes back to Kullback, Leibler [KL] and Ibragimov, Hasminskii [IH].

As we already mentioned Theorem 2.1 provides lower estimates for classes Θ with known lower estimates for the tight packing numbers $\bar{P}(\Theta,\delta)$. We now show how this theorem can be used in a situation when we know the behavior of packing numbers $P(\Theta,\delta)$.

LEMMA 2.1. *Let Θ be a compact subset of B. Assume that*

$$C_1\varphi(\delta) \le \ln P(\Theta,\delta) \le C_2\varphi(\delta), \quad \delta \in (0,\delta_1],$$

with a function $\varphi(\delta)$ satisfying the following condition. For any $\gamma > 0$ there is A_γ such that for any $\delta > 0$

$$\varphi(A_\gamma\delta) \le \gamma\varphi(\delta). \tag{2.3}$$

Then there exists $c_1 \ge 1$ and $\delta_2 > 0$ such that

$$\ln \bar{P}(\Theta,\delta,c_1,B) \ge C_3 \ln P(\Theta,\delta), \quad \delta \in (0,\delta_2].$$

Proof. For $\delta > 0$ we take the set $F := \{f_i\}_{i=1}^{P(\Theta,\delta)} \subset \Theta$ satisfying (2.1). Considering a $l\delta$-net with $l \ge 1$ for covering Θ we obtain that one of the balls of radius $l\delta$ contains at least $P(\Theta,\delta)/P(\Theta,l\delta)$ points of the set F. Denote this set of points by $F_l = \{f_i\}_{i\in\Lambda(l)}$. Then, obviously, for any $i \ne j \in \Lambda(l)$ we have

$$\delta \le \|f_i - f_j\| \le 2l\delta.$$

Therefore

$$\ln \bar{P}(\Theta, \delta, 2l, B) \geq \ln P(\Theta, \delta) - \ln P(\Theta, l\delta) \geq C_1\varphi(\delta) - C_2\varphi(l\delta).$$

Specifying $\gamma = C_1/(2C_2)$, $l = A_\gamma$, and $\delta_2 := \delta_1/l$ we continue

$$\geq C_1\varphi(\delta)/2 \geq \frac{C_1}{2C_2} \ln P(\Theta, \delta), \quad \delta \in (0, \delta_2].$$

As a corollary of Theorem 2.1 and Lemma 2.1 we obtain the following theorem.

THEOREM 2.2. *Assume Θ is a compact subset of $L_2(\mu)$ such that $\Theta \subset \frac{1}{4}U(\mathcal{C}(X))$ and*

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}. \tag{2.4}$$

Then there exist $\delta_0 > 0$ and $\eta_m := \eta_m(r) \asymp m^{-\frac{r}{1+2r}}$ such that

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \geq \delta_0 \quad \textit{for} \quad \eta \leq \eta_m \tag{2.5}$$

and

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \geq Ce^{-c(r)m\eta^2} \quad \textit{for} \quad \eta \geq \eta_m. \tag{2.6}$$

Proof. Condition (2.4) implies

$$C_1(r)\delta^{-1/r} \leq \ln P(\Theta, \delta) \leq C_2(r)\delta^{-1/r}, \quad \delta \in (0, \delta_1].$$

Clearly, the function $\varphi(\delta) = \delta^{-1/r}$ satisfies the condition (2.3) from Lemma 2.1. Therefore by Lemma 2.1 we obtain

$$\ln \bar{P}(\Theta, \eta, c_1(r), L_2(\mu)) \geq C_3(r)\eta^{-1/r}, \quad \eta \in (0, \delta_2(r)],$$

with some $c_1(r) \geq 1$. It remains to use Theorem 2.1 with η_m a solution of the equation

$$\frac{C_3(r)}{2}(2\eta)^{-1/r} - 8c_1(r)^2 m\eta^2 = 0.$$

It is clear that

$$\eta_m \asymp m^{-\frac{r}{1+2r}}.$$

REMARK 2.1. Theorem 2.2 holds in the case $\Theta \subset (M/4)U(\mathcal{C}(X))$, $|y| \leq M$, with constants allowed to depend on M.

We note that we do not impose direct restrictions on the measure μ in Theorem 2.2. However, the assumption (2.4) imposes an indirect restriction. For instance, if μ is a Dirac measure then we always have $\epsilon_n(\Theta, L_2(\mu)) \ll 2^{-n}$. Therefore, Theorem 2.2 does not apply in this case.

Let us make some comments on Theorem 2.2. It is clear that the parameter r controls the size of the compact Θ. The bigger the r the smaller the compact Θ. In the statement of Theorem 2.2 the parameter r affects the rate of decay of η_m. The quantity η_m is an important characteristic of the estimation process. The inequality (2.5) says that there is no way to estimate f_ρ from Θ with accuracy $\leq \eta_m$ with high confidence ($> 1 - \delta_0$). It seems natural that this critical accuracy η_m depends on the size of Θ (on parameter r). The inequalities (2.5) and (2.6) give

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \geq \delta_0 Ce^{-c(r)m\eta^2} \tag{2.7}$$

for all η. The exponent $m\eta^2$ in this inequality does not depend on the size of Θ. This may indicate that the form of this exponent is related not to the size of Θ but rather to the

stochastic nature of the problem. Other argument in support of the above observation is provided by an inequality from Section 5. We will use that inequality to show that in the case of a compact Θ consisting of only one function we have an analogue of (2.7) in the case of linear estimators. Let $\Theta = \{1/2\}$. Suppose that we are looking for a linear estimator

$$f_{\mathbf{z}} = \sum_{i=1}^{m} w_i(x_1, \dots, x_m, x) y_i \tag{2.8}$$

of the regression function f_ρ. Consider the following special case of the measure ρ. Let $\rho_X = \mu$ be any probabilistic measure on X. We define $\rho(y|x)$ as the Bernoulli measure:

$$\rho(1|x) = \rho(0|x) = 1/2, \quad x \in X.$$

Then for the above measure ρ we have $f_\rho(x) \equiv 1/2 \in \Theta$. Then

$$\|f_{\mathbf{z}} - f_\rho\|_{L_2(\mu)} \ge \int_X |f_{\mathbf{z}} - f_\rho| d\mu \ge \Big| \int_X (f_{\mathbf{z}} - f_\rho) d\mu \Big| = \Big| \sum_{i=1}^{m} w_i(x_1, \dots, x_m) y_i - 1/2 \Big|,$$

where

$$w_i(x_1, \dots, x_m) := \int_X w_i(x_1, \dots, x_m, x) d\mu.$$

Using Theorem 5.1 we get

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\mu)} \ge \eta\} \ge \mathrm{Prob}_{\mathbf{z} \in Z^m} \Big\{ \Big| \sum_{i=1}^{m} w_i(x_1, \dots, x_m) y_i - 1/2 \Big| \ge \eta \Big\}$$

$$\ge \exp(-25m\eta^2 - 6.25 \sum_{k=1}^{m-1} 1/k) \ge m^{-6.25} \exp(-25m\eta^2 - 1).$$

Therefore, in the case $\mathbb{E}(m)$ is the set of estimators of the form (2.8) we have for $\mathcal{M}_\mu := \{\rho : f_\rho = 1/2, \rho_X = \mu\}$

$$\mathbf{AC}_m(\mathcal{M}_\mu, \mathbb{E}, \eta) \ge m^{-6.25} \exp(-25m\eta^2 - 1).$$

We now proceed to upper estimates. In order to prove upper estimates we need to decide what should be the form of an estimator $f_{\mathbf{z}}$. In other words we need to specify the *hypothesis space* $\mathcal{H}$ (see [CS], [PS]) where an estimator $f_{\mathbf{z}}$ comes from.

The next question is how to build $f_{\mathbf{z}} \in \mathcal{H}$. In this paper we discuss a standard in statistics method of *empirical risk minimization* that takes

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

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*. We begin with the following estimate.

THEOREM 2.3 ([CS], [DKPT1,2]). *Assume that Θ satisfies (1.1). Suppose that $f_\rho \in \Theta$. Then for $\eta \ge A_0(M, D, r) m^{-\frac{r}{2(1+r)}}$*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z},\Theta} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le \exp(-c(M)m\eta^2). \tag{2.9}$$

Let us compare this theorem with Theorem 2.2. First of all we note that the estimator $E_{\mathbf{z}} : \mathbf{z} \to f_{\mathbf{z},\Theta}$ does not depend on η. Secondly, this estimator provides an optimal estimate for the probability distribution function with the exponent $m\eta^2$ that matches the exponent in the lower bound (2.6). However, (2.9) holds for $\eta \gg m^{-\frac{r}{2(1+r)}}$ and (2.6) holds for $\eta \gg m^{-\frac{r}{1+2r}}$. Thus Theorem 2.3 does not cover the range of $m^{-\frac{r}{1+2r}} \ll \eta \ll m^{-\frac{r}{2(1+r)}}$. Also, we should point out that Θ satisfies (1.1), which is stronger than the corresponding condition (1.3).

The key ingredient of the proof of Theorem 2.3 is the following theorem from [CS]. For a compact $\mathcal{H}$ denote

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

THEOREM 2.4 ([CS]). *Suppose that $\mathcal{H}$ is a compact subset of $\mathcal{C}(X)$ which is either convex or $f_\rho \in \mathcal{H}$. 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 $\epsilon > 0$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \epsilon\} \le N(\mathcal{H}, \epsilon/(24M), \mathcal{C}(X))2\exp\left(-\frac{m\epsilon}{288M^2}\right).$$

THEOREM 2.5 ([DKPT1,2]). *Let Θ satisfy (1.2). Suppose that $f_\rho \in \Theta$. Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \ge A_0(M,K,r)(\ln m/m)^{\frac{r}{1+2r}}$*

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

Theorem 2.5 allows us to build estimators with better accuracy than in Theorem 2.3: with error $\asymp (\ln m/m)^{\frac{r}{1+2r}}$ instead of error $\asymp m^{-\frac{r}{2(1+r)}}$. This is done under assumption (1.2) instead of (1.1). We note that condition (1.2) is stronger than (1.1). By Carl's inequality [C] (1.2) implies (1.1). We now describe the construction of the estimator $f_{\mathbf{z}}$ from Theorem 2.5. Let $\{L_n\}$ be a sequence of optimal (near optimal) subspaces for Θ, $\dim L_n = n$. Then for any $f \in \Theta$ there is a $\varphi_n \in L_n$ such that $\|f - \varphi_n\|_{\mathcal{C}(X)} \le 2Dn^{-r}$. It is clear that $\|\varphi_n\|_{\mathcal{C}(X)} \le 3D$. We now consider the set $V_n := 3DU(\mathcal{C}(X)) \cap L_n$. In other words we take as a hypothesis space the set V_n. We construct an estimator for $f_\rho \in \Theta$ by

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

with $n := [(\frac{m}{\ln m})^{\frac{1}{1+2r}}]$. This construction has an advantage over the choice $f_{\mathbf{z}} = f_{\mathbf{z},\Theta}$ in Theorem 2.3. Building $f_{\mathbf{z},V_n}$ we optimize over a ball in a finite dimensional space L_n instead of optimizing over Θ. We note that the set $\mathcal{H}$, smaller than Θ, that is used as a hypothesis space is known in statistics under the name *sieve* [G], [BM2]. In the proof of Theorem 2.5 we also use Theorem 2.4.

THEOREM 2.6 ([KT1]). *Let Θ satisfy (1.1). Suppose that $f_\rho \in \Theta$. Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \ge A_0(M,D,r)m^{-\frac{r}{1+2r}}$*

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

Comparing this theorem with Theorem 2.2 we see that Theorem 2.6 provides both the optimal rate of accuracy $\asymp m^{-\frac{r}{1+2r}}$ and the best estimate of probability distribution function with the exponent $m\eta^2$. The only thing in Theorem 2.6 that does not match the assumptions of Theorem 2.2 is the following. In Theorem 2.6 we assume that Θ satisfies (1.1) that means we impose restrictions in the uniform norm but not in the $L_2(\rho_X)$ norm

as in Theorem 2.2. Thus, Theorem 2.6 provides an optimal result in the case of Θ such that

$$\epsilon_n(\Theta, \mathcal{C}(X)) \asymp \epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}$$

for some measure μ.

The construction of $f_{\mathbf{z}}$ in Theorem 2.6 uses ϵ-nets of Θ in the uniform norm. We choose $\epsilon = A_0^{1/2} m^{-\frac{r}{1+2r}}$ and define V_ϵ to be a ϵ-net of Θ in the $\mathcal{C}(X)$ norm. We construct an estimator for $f_\rho \in \Theta$ by

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

The set V_ϵ is not convex and we cannot claim that $f_\rho \in V_\epsilon$. Therefore Theorem 2.4 does not apply for this set. In [KT1] we used the following theorem in the proof of Theorem 2.6.

THEOREM 2.7 ([DKPT1,2]). *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 $\epsilon > 0$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \epsilon\} \le N(\mathcal{H}, \epsilon/(24M), \mathcal{C}(X)) 2\exp\left(-\frac{m\epsilon}{C(M,R)}\right)$$

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

THEOREM 2.8 ([KT2]). *Assume that Θ satisfies (1.3) with $r > 1/2$. Suppose also $f_\rho \in \Theta$. Let $m\eta^4 \ge 1$. Then there exists an estimator $f_{\mathbf{z}}$ such that*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le C(M,r)\exp(-c(M)m\eta^4).$$

THEOREM 2.9 ([KT2]). *Let Θ satisfy (1.3). Suppose that $f_\rho \in \Theta$. Assume that $r \in (0, 1/2)$ and $m\eta^{2+1/r} \ge C_1(M, D, r)$. Then there exists an estimator $f_{\mathbf{z}}$ such that*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le C(M,D,r)\exp(-c(M,D,r)m\eta^{2+1/r}).$$

Assume that $r = 1/2$ and $m\eta^4/(1 + (\log(M/\eta))^2) \ge C_1(M, D)$. Then there exists an estimator $f_{\mathbf{z}}$ such that

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le C(M,D)\exp(-c(M,D)m\eta^4/(1 + (\log(M/\eta))^2)).$$

Theorems 2.8 and 2.9 are close to Theorem 2.2 in formulation of assumptions. In both cases we impose restrictions in the $L_2(\rho_X)$ norm. Combination of Theorems 2.2 and 2.9 gives the optimal rate of accuracy $\asymp m^{-\frac{r}{1+2r}}$ for classes $\mathcal{M}(\Theta, \mu)$ with

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}, \quad r \in (0, 1/2). \tag{2.10}$$

In the case $r > 1/2$ Theorems 2.2 and 2.8 do not match. It is an interesting open problem: find optimal rate of accuracy for classes $\mathcal{M}(\Theta, \mu)$ such that $\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}$ in the case $r > 1/2$.

The above discussed fact that in the case $r \in (0, 1/2)$ for any measure μ the behavior (2.10) of the entropy numbers determines the optimal rate of accuracy $\asymp m^{-\frac{r}{1+2r}}$ in the estimation problem indicates that it is natural to classify classes of priors by the behavior of their entropy numbers.

We now describe the construction of the estimator from Theorem 2.9. Contrary to the estimators from Theorems 2.3, 2.5, and 2.6 the estimator in Theorem 2.9 depends on η. Here we take $f_{\mathbf{z}} = f_{\mathbf{z},\mathcal{N}_\eta(\Theta)}$ with $\mathcal{N}_\eta(\Theta) := \mathcal{N}_\eta(\Theta, L_2(\rho_X))$. Proofs of Theorems 2.8 and 2.9 are somewhat more direct than the proofs of Theorems 2.3, 2.5, and 2.6. In the

proofs of Theorems 2.8 and 2.9 we use the Bernstein concentration measure inequality and apply the chaining technique (boot strapping technique, peeling device). We now formulate the Bernstein inequality. If ξ is a random variable (a real valued function on a probability space Z) then denote

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

The Bernstein inequality says: if $|\xi(z) - E(\xi)| \le M$ a.e. then for any $\epsilon > 0$

$$\text{Prob}_{z\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)\right| \ge \epsilon\right\} \le 2\exp\left(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}\right). \tag{2.11}$$

We complete the discussion of Theorem 2.9 by a theorem that is a corollary of Theorem 2.2, Remark 2.1, and Theorem 2.9.

THEOREM 2.10. *Let μ be a Borel measure on X. Assume $r \in (0, 1/2)$ and Θ is a compact subset of $L_2(\mu)$ such that*

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}, \qquad \Theta \subset (M/4)U(\mathcal{C}(X)).$$

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

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

and

$$C_1(\Theta, M)e^{-c_1(\Theta,M)m\eta^2} \le \mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \le C_2(\Theta, M)e^{-c_2(\Theta,M)m\eta^{2+1/r}}$$

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

The above theorems give the upper estimates in the following style. For a given class $\mathcal{M}$ there exist $\eta_m^+(\mathcal{M})$ and positive constants C, c, a such that for $\eta \ge \eta_m^+(\mathcal{M})$

$$\mathbf{AC}_m(\mathcal{M}, \eta) \le Ce^{-cm\eta^a}.$$

Theorem 2.1 and 2.2 give the lower estimates of the following type. For a given $\mathcal{M}$ there exist $\delta_0(\mathcal{M}) > 0$ and $\eta_m^-(\mathcal{M}) > 0$ such that for $\eta \le \eta_m^-(\mathcal{M})$ one has

$$\mathbf{AC}_m(\mathcal{M}, \eta) \ge \delta_0(\mathcal{M}).$$

These inequalities indicate that the behavior of the accuracy confidence function changes dramatically within the *critical interval* $[\eta_m^-(\mathcal{M}), \eta_m^+(\mathcal{M})]$. It drops from a constant $\delta_0(\mathcal{M})$ to an exponentially small quantity $C\exp(-cm\eta_m^+(\mathcal{M})^a)$. One may also call the interval $[\eta_m^-(\mathcal{M}), \eta_m^+(\mathcal{M})]$ *the interval of phase transition*. Clearly, good estimates for $\eta_m^-(\mathcal{M})$ and $\eta_m^+(\mathcal{M})$ are of great importance. We introduce more terminology in this regard. Suppose for a given class $\mathcal{M}$ there exist a function $\varphi(\mathcal{M}, m)$ and two constants $C_1(\mathcal{M})$, $C_2(\mathcal{M})$ such that

$$C_1(\mathcal{M})\varphi(\mathcal{M}, m) \le \eta_m^-(\mathcal{M}) \le \eta_m^+(\mathcal{M}) \le C_2(\mathcal{M})\varphi(\mathcal{M}, m).$$

Then we call the function $\varphi(\mathcal{M}, m)$ *the critical rate of accuracy*. The following theorem is a corollary of Theorem 2.10.

THEOREM 2.11. *Let $r \in (0, 1/2)$. Assume Θ is a compact subset of $L_2(\mu)$ such that*

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}, \qquad \Theta \subset (M/4)U(\mathcal{C}(X)).$$

Let $\mathcal{M}(\Theta,\mu) := \{\rho : f_\rho \in \Theta, \rho_X = \mu, |y| \le M\}$. *Then the critical rate of accuracy exists for* $\mathcal{M}(\Theta,\mu)$ *and has the order*

$$\varphi(\mathcal{M}(\Theta,\mu),m) \asymp m^{-\frac{r}{1+2r}}.$$

Results of this section show that from a theoretical point of view the entropy numbers $\epsilon_n(\Theta, L_2(\rho_X))$ are the right characteristic of a class Θ in the problem of estimating the regression function f_ρ. However, the above discussion indicates certain difficulties with the use of the entropy numbers $\epsilon_n(\Theta, L_2(\rho_X))$. As we have mentioned the estimator $f_{\mathbf{z}}$ from Theorem 2.9 has been built using the η-net of Θ in the $L_2(\rho_X)$ norm. In many cases the measure ρ_X is unknown. Therefore, we would like to construct an estimator that does not depend on ρ_X and provides good estimation for all ρ_X. This is the main goal of distribution-free theory of regression. One of the ways out of the above problem with the use of the characteristic $\epsilon_n(\Theta, L_2(\rho_X))$ is to go through the uniform norm, i.e. to use the characteristic $\epsilon_n(\Theta, \mathcal{C}(X))$. Clearly, this narrows the set of classes of priors Θ we can work with. Theorem 2.6 shows that we can construct an estimator $f_{\mathbf{z}}$ that does not depend on ρ_X and does an optimal (in the sense of order) job for classes satisfying (1.1). From a theoretical point of veiw this estimator is very good. However, it is clear that we have a problem with direct practical implementation of this estimator because it is built on the base of an ϵ-net of Θ. The estimator from Theorem 2.5 is better in the sense of implementation. It is constructed by least squares method in the finite dimensional subspace L_n. Thus in addition to theoretical problem of finding optimal rates of estimation we have a practical problem of implementation of optimal (near optimal) estimators. We want to understand what characteristics of prior classes Θ are suitable for the task of convenient practical implementation. It is somewhat clear that the description of Θ in terms of the entropy numbers does not fit this goal. Indeed, at this point it looks unfeasible to implement algorithms based on ϵ-nets of function classes.

Interesting results in this direction on building estimation schemes with nice implementation properties have been obtained in the recent paper [BCDDT]. The most important property of those estimation schemes is universality. It is a very important property of an estimation algorithm. We do not discuss the universality property in this paper and refer the reader to the papers [DKPT1], [DKPT2], [BCDDT], [KT2] where this property has been discussed in detail.

We present here a result from [KT2] in a style of Theorem 2.5 with a description of Θ in the $L_2(\rho_X)$ norm instead of the $\mathcal{C}(X)$ norm. Let $B(X)$ be a Banach space with the norm $\|f\|_{B(X)} := \sup_{x\in X} |f(x)|$. Let $\{L_n\}_{n=1}^\infty$ be a given sequence of n-dimensional linear subspaces of $B(X)$ such that L_n is also a subspace of each $L_\infty(\mu)$, where μ is a probability measure on X, $n = 1, 2, \dots$. Assume that n-dimensional linear subspaces L_n have the following property: for any probability measure μ on X one has

$$\|P_{L_n}^\mu\|_{B(X)\to B(X)} \le K, \quad n = 1, 2, \dots \tag{2.12}$$

where P_L^μ is the operator of $L_2(\mu)$ projection onto L. For a finite dimensional linear subspace $L \subset L_2(\rho_X)$ and $f \in L_2(\rho_X)$ we denote by $d(f,L)_{L_2(\rho_X)}$ the $L_2(\rho_X)$ distance between f and L.

THEOREM 2.12 ([KT2]). *Assume that a sequence $\{L_n\}_{n=1}^{\infty}$ satisfies* (2.12). *For given m, $r > 0$ there exists an estimator $f_{\mathbf{z}}$ such that for any ρ satisfying*

$$d(f_\rho, L_n)_{L_2(\rho_X)} \le Dn^{-r}, \quad n = 1, 2, \dots,$$

we get for $\eta \ge A_0(M, K, r)(\ln m/m)^{\frac{r}{1+2r}}$

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

The above theorem can be used, in particular, in the following situation. Let X be a compact subset of $\mathbb{R}^d$. Let $\mathcal{P}_n$ denote the set of all partitions of X into n disjoint Borel subsets. Let $p_n \in \mathcal{P}_n$, $n = 1, \dots$. Define L_n as a subspace of all functions that are piecewise constant on the partition p_n. The subspaces L_n satisfy (2.12) with $K = 1$.

Thus we can obtain simpler estimators when we replace assumptions on Θ in terms of entropy numbers (a characteristic of nonlinear approximation) by assumptions on Θ in terms of approximation by linear subspaces (a characteristic of linear approximation). It is known from works in approximation theory (see surveys [D], [T]) and statistics ([DJ], [KP]) that nonlinear approximation is more flexible than linear approximation and provides optimal means of approximation and estimation. The most important in this regard form of nonlinear approximation is the n-term approximation with regard to a given basis or more generally with regard to a dictionary. We present one result in this direction from [DKPT1]. We will consider n-term approximations with regard to a given system Ψ. Assume that the system $\Psi = \{\psi_j\}_{j=1}^{\infty}$ is a (VP)-system, i.e. satisfies the condition:

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

$$P_k(\psi_j) = \lambda_{k,j}\psi_j,$$
$$\lambda_{k,j} = 1 \quad \text{for} \quad j = 1, \dots, n_k; \qquad \lambda_{k,j} = 0 \quad \text{for} \quad j > A_2 n_k,$$
$$\|P_k\|_{\mathcal{C}(X)\to\mathcal{C}(X)} \le A_3, \quad k = 1, 2, \dots.$$

Denote

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

and

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

THEOREM 2.13. *Let $f_\rho \in \Theta$ and let Θ satisfy the following two conditions.*

$$\sigma_n(\Theta, \Psi) \le C_1 n^{-r}, \quad \Theta \subset C_1 U(\mathcal{C}(X)),$$
$$E_n(\Theta, \Psi) := \sup_{f\in\Theta} \inf_{c_1,\dots,c_n} \|f - \sum_{j=1}^{n} c_j\psi_j\|_{\mathcal{C}(X)} \le C_2 n^{-b},$$

where Ψ is the (VP)-system. Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \ge A_0(M, r, b)(\ln m/m)^{\frac{r}{1+2r}}$

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

We note that the trigonometric system and wavelets are (VP)-systems.

We now give a concrete example of a class of priors Θ to demonstrate how the general theory developed in this section works. Let $X = [0,1]^d$ and W_p^s, $s \in \mathbb{N}$, $1 \le p \le \infty$, be the Sobolev class (the unit ball of the Sobolev space): the set of all functions $g \in L_p(X)$ whose distributional derivatives $D^\nu g$, $\|\nu\|_{\ell_1} \le s$, are also in $L_p(X)$ and

$$\sum_{\|\nu\|_{\ell_1} \le s} \|D^\nu g\|_{L_p(X)} \le 1.$$

Then it is known [BS] that for $s > d/p$ one has

$$\epsilon_n(W_p^s, \mathcal{C}) \asymp n^{-r}, \quad r := s/d,$$

and

$$\epsilon_n(W_p^s, L_2) \asymp n^{-r}.$$

Then by Theorem 2.6

$$\mathbf{AC}_m(W_p^s, \eta) \le e^{-c_1(M)m\eta^2}, \quad \eta \ge \eta_m^+ \asymp m^{-\frac{r}{1+2r}}.$$

By Theorem 2.2 and Remark 2.1 with μ - Lebesgue measure we get

$$\mathbf{AC}_m(W_p^s, \eta) \ge \delta_0, \quad \eta \le \eta_m^- \asymp m^{-\frac{r}{1+2r}},$$
$$\mathbf{AC}_m(W_p^s, \eta) \ge Ce^{-c_2(M)m\eta^2}, \quad \eta \ge \eta_m^-.$$

These results give a very accurate description of the accuracy confidence function $\mathbf{AC}_m(W_p^s, \eta)$.

We complete this section by a remark concerning the quantities $E(\Theta, m)$ that give the rate of accuracy of optimal estimation in the sense of expectation. We have already mentioned in the Introduction (see (1.4), (1.5)) how the accuracy confidence function $\mathbf{AC}_m(\Theta, \eta)$ can be used for estimating $E(\Theta, m)$ from below and from above. We now develop the ideas of (1.4) and (1.5) to obtain the right order of

$$E(\Theta, m)_q := \inf_{\hat{f}} \sup_{f_\rho \in \Theta} E_{\rho^m}(\|f_\rho - \hat{f}\|_{L_2(\rho_X)}^q), \quad 0 < q < \infty.$$

Suppose that a class Θ is such that there exists a critical rate $\varphi(\Theta, m) := \varphi(\mathcal{M}(\Theta), m)$ of accuracy for this class and for any $q \in (0, \infty)$ we have $\mathbf{AC}_m(\Theta, \eta_m^+) \ll \varphi(\Theta, m)^q$. Then on one hand for any $f_{\mathbf{z}}$

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)}^q) \ge \int_0^\infty \mathbf{AC}_m(\Theta, \eta^{1/q}) d\eta \ge \delta_0(\eta_m^-)^q \gg \varphi(\Theta, m)^q.$$

On the other hand for $\eta = \eta_m^+$ there exists $f_{\mathbf{z}}$ such that

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)}^q) \le (\eta_m^+)^q + (2M)^q \mathbf{AC}_m(\Theta, \eta_m^+) \ll \varphi(\Theta, m)^q.$$

In particular, this implies that for any $0 < q < \infty$ we have for $1 \le p \le \infty$, $s > d/p$

$$E(W_p^s, m)_q \asymp m^{-\frac{qr}{1+2r}}, \quad r := s/d. \tag{2.13}$$

In the case $q = 2$ the lower estimate in (2.13) has been obtained by Stone [S] in 1982. The corresponding upper estimate and a discussion can be found in [GKKW].

3. No prior on f_ρ. In this section we briefly discuss the following setting. We now do not impose any restriction on the unknown measure ρ, except our standard assumption $|y| \le M$. In such a situation we, clearly, cannot estimate f_ρ with a nontrivial error estimate. Instead of estimating f_ρ we now estimate the $L_2(\rho_X)$ projection of f_ρ onto a compact W that we may choose. This setting is a more general setting than the one from Section 2. Indeed, if we know that $f_\rho \in \Theta$ then $f_\Theta = (f_\rho)_\Theta = f_\rho$. Therefore, the results of this section apply with $W = \Theta$. This remark motivates us to impose restrictions on W in the same style as we did in Section 2. We begin with the upper estimates. For a compact in $L_2(\rho_X)$ set W denote by $f_W := (f_\rho)_W$ the $L_2(\rho_X)$-projection of f_ρ onto W. In other words

$$f_W := \arg\min_{f\in W} \mathcal{E}(f).$$

Let us denote

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

THEOREM 3.1 ([CS], [DKPT1]). *Assume that $W \in \mathcal{S}^r$. Then for $\eta \ge A_0(M, D, r)m^{-\frac{r}{2(1+2r)}}$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \ge \eta^2\} \le \exp(-c(M)m\eta^4).$$

THEOREM 3.2 ([KT1]). *Assume that W satisfies (1.1). Then we have the following estimates*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \ge \eta^2\} \le C(M, D, r)\exp(-c(M)m\eta^4),$$

provided $r > 1/2$, $m\eta^4 \ge 1$;

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \ge \eta^2\} \le C_1(M, D)\exp(-c(M, D)m\eta^4/(1 + (\log(M/\eta))^2)),$$

provided $r = 1/2$, $m\eta^4/(1 + (\log(M/\eta))^2) \ge C_2(M, D)$;

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \ge \eta^2\} \le C_1(M, D, r)\exp(-c(M, D, r)m\eta^{2/r}),$$

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

In Theorems 3.1 and 3.2 we choose the $f_{\mathbf{z},W}$ as the estimator. Theorem 3.2 gives the following upper estimate for the accuracy confidence function. For $W \in \mathcal{S}^r$, $r > 1/2$ we have

$$\mathbf{AC}^p_m(W, \eta) \le C(M, D, r)\exp(-c(M)m\eta^4) \quad \text{for} \quad \eta \ge m^{-1/4}. \tag{3.1}$$

Let us compare this estimate with the corresponding estimate for $\mathbf{AC}_m(\Theta, \eta)$. Theorem 2.6 gives for $\Theta \in \mathcal{S}^r$

$$\mathbf{AC}_m(\Theta, \eta) \le \exp(-c(M)m\eta^2) \quad \text{for} \quad \eta \gg m^{-\frac{r}{1+2r}}. \tag{3.2}$$

The estimates (3.1) and (3.2) differ in two ways. First, the accuracy $\asymp m^{-\frac{r}{1+2r}}$ in (3.2) depends on r and better for $r > 1/2$ than the accuracy $\asymp m^{-1/4}$ in (3.1) that does not depend on r. Second, the exponent $m\eta^2$ from (3.2) in the bound for the probability distribution function is better than the corresponding exponent $m\eta^4$ from (3.1). The following proposition shows that we cannot improve (3.1).

PROPOSITION 3.1. *There exist two positive constants c_1, c_2 and a class W consisting of two functions 1 and -1 such that for every $m = 2, 3, \dots$ and $m^{-1/4} \le \eta \le 1/2$ there are*

two measures ρ_0 and ρ_1 such that for any estimator $f_{\mathbf{z}} \in W$ for one of $\rho = \rho_0$ or $\rho = \rho_1$ we have

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_W) \geq \eta^2\} \geq c_1 \exp(-c_2 m\eta^4).$$

In the case $\eta = m^{-1/4}$ this proposition has been proved in [KT1]. The proof in the general case $m^{-1/4} \leq \eta \leq 1/2$ is similar. Proposition 3.1 indicates that there is a phenomenon of saturation for collections $\mathcal{S}^r$ for $r > 1/2$.

In the case $r \in (0, 1/2)$ Theorem 3.2 gives the estimate

$$\mathbf{AC}_m^p(W, \eta) \ll \exp(-c(M, D, r)m\eta^{2/r}) \quad \text{for} \quad \eta \gg m^{-r/2}. \tag{3.3}$$

Similarly to the above comparison of (3.1) and (3.2) we see that (3.3) is weaker than (3.2). The following proposition from [KT1] shows that the accuracy bound in (3.3) cannot be improved on the whole collection $\mathcal{S}^r$.

PROPOSITION 3.2 ([KT1]). *For any $r \in [0, 1/2]$ and for every $m \in \mathbb{N}$ there is $W \subset U(L_\infty([0,1])$ satisfying $\epsilon_n(W, L_\infty) \leq n^{-r}$ for $n \in \mathbb{N}$ such that for every estimator $f_{\mathbf{z}} \in W$ there is a ρ such that*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}((f_\rho)_W) \geq m^{-r}/4\} \geq 1/7.$$

We now present two results in the case of W satisfying a weaker condition (1.3) instead of (1.1).

THEOREM 3.3 ([KT2]). *Assume that W satisfies (1.3) with $r > 1/2$. Let $m\eta^{2(1+\max(1/r,1))} \geq A_0(M, D, r) \geq 1$. Then there exists an estimator $f_{\mathbf{z}} \in W$ such that*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_W) \geq \eta^2\} \leq C_1(M, D, r) \exp(-c_1(M)m\eta^4).$$

THEOREM 3.4 ([KT2]). *Assume that W satisfies (1.3) with $r \in (0, 1/2)$. Let $m\eta^{2(1+1/r)} \geq A_0(M, D, r) \geq 1$. Then there exists an estimator $f_{\mathbf{z}} \in W$ such that*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_W) \geq \eta^2\} \leq C(M, D, r) \exp(-c(M, D, r)m\eta^{2+1/r}).$$

We now give an idea of proofs of the upper estimates of this section. This idea provides a motivation for our interest in the problem discussed in the next section. Let W be a hypothesis space. Then we have

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) &= \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},W}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},W}) - \mathcal{E}_{\mathbf{z}}(f_W) + \mathcal{E}_{\mathbf{z}}(f_W) - \mathcal{E}(f_W) \\ &\leq \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},W}) + \mathcal{E}_{\mathbf{z}}(f_W) - \mathcal{E}(f_W).\end{aligned}$$

Thus we want to estimate

$$\sup_{f \in W} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)|.$$

4. Estimates for $L_{\mathbf{z}}(f)$. One of important questions discussed in [CS], [DKPT1], [DKPT2], [KT1], [KT2] is to estimate the *defect function* $L_{\mathbf{z}}(f) := L_{\mathbf{z},\rho}(f) := \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)$ of $f \in W$. If ξ is a random variable (a real valued function on a probability space Z) then denote as above

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

In this section it will be convenient for us to assume that

$$\text{for all} \quad f \in W, \quad f : X \to Y \quad \text{is such that} \quad |f(x) - y| \leq M \quad \text{a.e.} \tag{4.1}$$

THEOREM 4.1 ([CS]). *Let W be a compact subset of $\mathcal{C} := \mathcal{C}(X)$. Assume that ρ, W satisfy* (4.1). *Then, for all $\eta > 0$*

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \le N(W, \eta/(16M), \mathcal{C}) 2 \exp\left(-\frac{m\eta^2}{8(4\sigma^2 + M^2\eta/3)}\right). \tag{4.2}$$

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

REMARK 4.1. In general we cannot guarantee that the set $\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\}$ is ρ^m-measurable. In such a case the relation (4.2) and further relations of this type are understood in the sense of outer measure associated with the ρ^m. For instance, for (4.2) this means that there exists ρ^m-measurable set G such that $\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \subset G$ and (4.2) holds for G.

In [CS] this theorem has been derived from Bernstein's inequality (2.11). We note that other variants of this theorem can be found in the literature (see, for instance, [Po], [GKKW]). Theorem 4.1 contains a factor $N(W, \eta/(16M), \mathcal{C})$ that may grow exponentially for classes W satisfying (1.1): $N(W, \eta, \mathcal{C}) \le 2^{(D/\eta)^{1/r}+1}$. A stronger (in a certain sense) estimate than (4.2) has been obtained in [KT1] under the assumption that W satisfies (1.1).

THEOREM 4.2 ([KT1]). *Assume that ρ, W satisfy* (4.1) *and W is such that*

$$\sum_{n=1}^{\infty} n^{-1/2} \epsilon_n(W, \mathcal{C}) < \infty. \tag{4.3}$$

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

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M, \epsilon(W)) \exp(-c(M) m\eta^2)$$

with $C(M, \epsilon(W))$ that may depend on M and $\epsilon(W) := \{\epsilon_n(W, \mathcal{C})\}$; $c(M)$ may depend only on M.

THEOREM 4.3 ([KT1]). *Assume that ρ, W satisfy* (4.1) *and W is such that*

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

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

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

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

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M, \epsilon(W)) \exp(-c(M) m\eta^2/S_J^2).$$

We formulate two corollaries of Theorem 4.3.

COROLLARY 4.1 ([KT1]). *Assume ρ, W satisfy* (4.1) *and $\epsilon_n(W, \mathcal{C}) \le Dn^{-r}$, $r \in (0, 1/2)$. Then for m, η satisfying $m\eta^{1/r} \ge C_1(M, D, r)$ we have*

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M, D, r) \exp(-c(M, D, r) m\eta^{1/r}).$$

COROLLARY 4.2 ([KT1]). *Assume ρ, W satisfy (4.1) and $\epsilon_n(W,\mathcal{C}) \le Dn^{-r}$, $r \in (0,1/2)$. Then for m, η, $\delta \ge \eta/(8M)$ satisfying $m\eta^2\delta^{1/r-2} \ge C_1(M,D,r)$ we have*

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{N}_\delta(W)} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M,D,r)\exp(-c(M,D,r)m\eta^2\delta^{1/r-2})$$

where $\mathcal{N}_\delta(W)$ is a minimal δ-net of W in the $\mathcal{C}$ norm.

In [KT2] we have proved that it is impossible to have even a weaker form of Theorem 4.2 if we use the $L_2(\rho_X)$ norm instead of the uniform norm $\mathcal{C}$. However, it turned out that we can prove an $L_2(\rho_X)$ analogue of Theorem 4.2 for the δ-net $\mathcal{N}_\delta(W)$ of W in the $L_2(\rho_X)$ norm instead of W for $\delta^2 \ge \eta$. The following proposition shows that if we consider entropy of W in $L_2[0,1)$ rather than in $\mathcal{C}[0,1]$ then even a fast decay of $\epsilon_n(W,L_2(\rho_X))$ (say, $\epsilon_n(W,L_2(\rho_X)) = o(n^{-r})$ for every $r > 0$) does not guarantee nontrivial estimates for $\sup_{f\in W}|L_{\mathbf{z}}(f)|$. We assume that $Y = [-1,1]$, and thus, the functions $f \in W$ and f_ρ are uniformly bounded.

PROPOSITION 4.1 ([KT2]). *Let N be a non-increasing mapping $(0,+\infty) \to [1,+\infty)$ such that*

$$\lim_{u\to 0+} \log N(u)/\log(1/u) = +\infty.$$

Then there exist a set $W \subset U(L_\infty[0,1))$ and a ρ such that

$$N(W,\epsilon,L_2(\rho_X)) \le N(\epsilon)$$

and for every m

$$\rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z}}(f)| \le 1/2\} = 0.$$

THEOREM 4.4 ([KT2]). *Assume that ρ, W satisfy (4.1) and W is such that*

$$\sum_{n=1}^{\infty} n^{-1/2}\epsilon_n(W,L_2(\rho_X)) < \infty.$$

Let $m\eta^2 \ge 1$. Then for any δ satisfying $\delta^2 \ge \eta$ we have for a minimal δ-net $\mathcal{N}_\delta(W)$ of W in the $L_2(\rho_X)$ norm

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{N}_\delta(W)} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M,\epsilon(W))\exp(-c(M)m\eta^2).$$

THEOREM 4.5 ([KT2]). *Assume that ρ, W satisfy (4.1) and*

$$\sum_{n=1}^{\infty} n^{-1/2}\epsilon_n = \infty, \quad \epsilon_n := \epsilon_n(W,L_2(\rho_X)).$$

Let η,δ be such that $\delta^2 \ge \eta$. Define $J := J(\delta)$ as the minimal j satisfying $\epsilon_{2^j} \le \delta$ and

$$S_J := \sum_{j=1}^{J} 2^{(j+1)/2}\epsilon_{2^{j-1}}, \quad J \ge 1; \quad S_0 := 1.$$

Then for m, η satisfying $m(\eta/S_J)^2 \ge 36M^2$ we have

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{N}_\delta(W)} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M,\epsilon(W))\exp(-c(M)m(\eta/S_J)^2),$$

where $\mathcal{N}_\delta(W)$ is a minimal δ-net of W in the $L_2(\rho_X)$.

COROLLARY 4.3 ([KT2]). *Assume ρ, W satisfy (4.1) and $\epsilon_n(W, L_2(\rho_X)) \le Dn^{-r}$, $r \in (0, 1/2)$. Then for m, η, $\delta^2 \ge \eta$ satisfying $m\eta^2\delta^{1/r-2} \ge C_1(M, D, r)$ we have*

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{N}_\delta(W)} |L_{\mathbf{z}}(f)| \ge \eta\} \le C(M, D, r)\exp(-c(M, D, r)m\eta^2\delta^{1/r-2}),$$

where $\mathcal{N}_\delta(W)$ is a minimal δ-net of W in the $L_2(\rho_X)$.

On the base of the above discussion we propose to study the following function that we call the *accuracy confidence function for the defect function*. Let a function class W and a set $\mathcal{M}$ of admissible measures ρ be given. For $m \in \mathbb{N}$, $\eta > 0$ we define

$$\mathbf{AC}_m^d(W, \mathcal{M}, \eta) := \sup_{\rho\in\mathcal{M}} \rho^m\{\mathbf{z} : \sup_{f\in W} |L_{\mathbf{z},\rho}(f)| \ge \eta\}.$$

We note that the above function is related to the concept of the Glivenko-Cantelli sample complexity of a class Φ with accuracy η and confidence δ:

$$S_\Phi(\epsilon, \delta) := \min\Big\{n : \forall m \ge n,\ \forall\rho \\ \rho^m\Big\{\mathbf{z} = (z_1, \dots, z_m) : \sup_{\phi\in\Phi}\Big|\int_Z \phi d\rho - \frac{1}{m}\sum_{i=1}^m \phi(z_i)\Big| \ge \eta\Big\} \le \delta\Big\}.$$

In order to see that we define $z_i := (x_i, y_i)$, $i = 1, \dots, m$; $\phi(x, y) := (f(x) - y)^2$; $\Phi := \{(f(x) - y)^2, f \in W\}$. One can find a survey of recent results on the Glivenko-Cantelli sample complexity in [M].

Theorem 4.2 asserts that for W satisfying (4.3) and for $\mathcal{M}$ satisfying (4.1) we have

$$\mathbf{AC}_m^d(W, \mathcal{M}, \eta) \le C(M, \epsilon(W))\exp(-c(M)m\eta^2), \quad \eta \ge m^{-1/2}.$$

Corollary 4.1 says that for W satisfying (1.1) with $r \in (0, 1/2)$ and for $\mathcal{M}$ satisfying (4.1) we have

$$\mathbf{AC}_m^d(W, \mathcal{M}, \eta) \le C(M, D, r)\exp(-c(M, D, r)m\eta^{1/r}), \quad \eta \gg m^{-r}.$$

It turns out that in some applications it is more convenient to have an estimate of the $\mathbf{AC}_m^d$-function for a minimal δ-net of W instead of W itself. Corollary 4.2 gives the following estimate under the assumption that W satisfies (1.1) with $r \in (0, 1/2)$, $\mathcal{M}$ satisfies (4.1) and $\delta \ge \eta/(8M)$:

$$\mathbf{AC}_m^d(\mathcal{N}_\delta(W, \mathcal{C}), \mathcal{M}, \eta) \le C(M, D, r)\exp(-c(M, D, r)m\eta^2\delta^{1/r-2}).$$

Let now μ be a fixed probability measure on X. Assume W is such that

$$\sum_{n=1}^\infty n^{-1/2}\epsilon_n(W, L_2(\mu)) < \infty.$$

Consider $\mathcal{M}(W, \mu) := \{\rho \quad \text{satisfying} \quad (4.1) : \rho_X = \mu\}$. Then Theorem 4.4 claims that for any μ we have for $\delta^2 \ge \eta \ge m^{-1/2}$

$$\mathbf{AC}_m^d(\mathcal{N}_\delta(W, L_2(\mu)), \mathcal{M}(W, \mu), \eta) \le C(M, \epsilon(W))\exp(-c(M)m\eta^2).$$

Corollary 4.3 states that for W satisfying $\epsilon_n(W, L_2(\mu)) \le Dn^{-r}$, $r \in (0, 1/2)$ we have for $\delta^2 \ge \eta$

$$\mathbf{AC}_m^d(\mathcal{N}_\delta(W, L_2(\mu)), \mathcal{M}(W, \mu), \eta) \le C(M, D, r)\exp(-c(M, D, r)m\eta^2\delta^{1/r-2}).$$

5. Lower estimates for the Bernoulli scheme. We consider in this section the following estimation problem. Let y be a random variable such that

$$\mathrm{Prob}\{y=1\}=\mathrm{Prob}\{y=0\}=1/2.$$

Then $E(y)=1/2$. We begin our discussion with the standard estimator $f_m:=m^{-1}\sum_{i=1}^m y_i$. Then it is well known that

$$\mathrm{Prob}\{|f_m-1/2|\ge\epsilon\}=2^{-m}\Big(\sum_{|k-m/2|\ge m\epsilon} C_m^k\Big),$$

where C_m^k are the binomial coefficients. It is easy to check that

$$C_1 e^{-c_1 m\epsilon^2}\le\sum_{|k-m/2|\ge m\epsilon} C_m^k\le C_2 e^{-c_2 m\epsilon^2}$$

with positive absolute constants C_1, C_2, c_1, c_2.

The main goal of this section is to prove that f_m is optimal in a certain sense among all linear estimators. We will prove the following theorem.

THEOREM 5.1. *For any $\epsilon\in[0,1/2]$, $m\ge 2$, and $w=(w_1,\dots,w_m)$ we have*

$$\mathrm{Prob}\Big\{\Big|\sum_{i=1}^m w_i y_i-1/2\Big|\ge\epsilon\Big\}\ge\exp\Big(-cm\epsilon^2-\frac{c}{4}\sum_{k=1}^{m-1}\frac{1}{k}\Big)$$

with $c=25$.

We begin with a technical lemma.

LEMMA 5.1. *Let $\epsilon\in(0,\beta]$, $9n\ge\epsilon^{-2}$, $w_n\in[0,1/n]$. Then for $\epsilon_1:=(\epsilon-w_n/2)(1-w_n)^{-1}$, $\epsilon_2:=(\epsilon+w_n/2)(1-w_n)^{-1}$ one has for $c=25$, $\beta=(\ln 2)^{1/2}/5$*

$$\exp(-c(n-1)\epsilon_1^2)+\exp(-c(n-1)\epsilon_2^2)\ge 2\exp\Big(-cn\epsilon^2-\frac{c}{4(n-1)}\Big). \tag{5.1}$$

Proof. We consider separately two cases: I $w_n\in[0,1/(2n)]$ and II $w_n\in(1/(2n),1/n]$.

CASE I. Using the convexity of function e^{-x} we obtain for any $C>0$

$$\exp(-C(n-1)\epsilon_1^2)+\exp(-C(n-1)\epsilon_2^2)\ge 2\exp(-C(n-1)(\epsilon_1^2+\epsilon_2^2)/2). \tag{5.2}$$

Next,

$$\epsilon_1^2+\epsilon_2^2=(1-w_n)^{-2}((\epsilon-w_n/2)^2+(\epsilon+w_n/2)^2)=(1-w_n)^{-2}(2\epsilon^2+w_n^2/2).$$

Using the inequality

$$\frac{n-1}{(1-w_n)^2}\le n\quad\text{for}\quad w_n\in[0,1/(2n)]$$

we get

$$(n-1)(\epsilon_1^2+\epsilon_2^2)/2\le n\epsilon^2+1/(16n). \tag{5.3}$$

Substituting (5.3) into (5.2) we obtain (5.1).

CASE II. We rewrite

$$\begin{aligned}S&:=\exp(-c(n-1)\epsilon_1^2)+\exp(-c(n-1)\epsilon_2^2)\\&=\exp(-c(n-1)\epsilon_1^2)(1+\exp(-c(n-1)(\epsilon_2^2-\epsilon_1^2))).\end{aligned}$$

We have an identity

$$\epsilon_2^2 - \epsilon_1^2 = 2w_n\epsilon(1-w_n)^{-2}.$$

Denote $a_n := (n-1)(1-w_n)^{-2}$. We have

$$1 - 1/n \le a_n/n \le n/(n-1). \tag{5.4}$$

Let us estimate $\delta := n\epsilon^2 - (n-1)\epsilon_1^2$. We have

$$\delta = \epsilon^2\left(\frac{n}{n-1}(1-w_n)^2 - 1\right)a_n + a_n w_n\epsilon - a_n w_n^2/4.$$

Using

$$\frac{n}{n-1}(1-w_n)^2 - 1 = \frac{(1-w_n)^2}{1-1/n} - 1 \ge 1 - w_n - 1 = -w_n$$

we get

$$\delta \ge a_n w_n\epsilon - a_n w_n\epsilon^2 - a_n w_n^2/4.$$

Therefore

$$S \ge \exp(-cn\epsilon^2 - ca_n w_n^2/4)2\cosh(ca_n w_n\epsilon)\exp(-ca_n w_n\epsilon^2).$$

We note that by (5.4)

$$a_n w_n^2 \le a_n n^{-2} \le (n-1)^{-1}.$$

Thus we proceed to estimating $\cosh(A\epsilon)\exp(-A\epsilon^2)$ with $A := ca_n w_n$. By (5.4) and by our assumption $w_n > 1/(2n)$ we get

$$A \ge c(1-1/n)/2 \ge c/3, \quad n = 3, \dots. \tag{5.5}$$

It is easy to check that for the function $f(x) := \cosh(Ax) - \exp(Ax^2)$ we have $f(0) = 0$ and $f'(x) \ge 0$ for $x^2 \le (\ln 4)/A$ in the case $A \ge 8$. The latter inequality $A \ge 8$ follows from (5.5). Therefore,

$$\cosh(A\epsilon)\exp(-A\epsilon^2) \ge 1 \quad \text{if} \quad \epsilon^2 \le \ln 4/A.$$

By (5.4) we have $A \le cn/(n-1)$ and, hence, for $c = 25$ and $n \ge 2$ we have $\beta^2 = (1/5)^2\ln 2 \le \ln 4/A$ for all A of the form $A = ca_n w_n$. This completes the proof of the lemma.

LEMMA 5.2. *For any $\epsilon \in [0, 1/2]$, $m \ge 2$, and $w_1 \ge w_2 \ge \cdots \ge w_m \ge 0$, $\sum_{i=1}^m w_i = 1$ we have*

$$|\{\Lambda \subseteq [1,m] : \sum_{i\in\Lambda} w_i \ge 1/2 + \epsilon\}| \ge 2^m \exp\left(-cm\epsilon^2 - \frac{c}{4}\sum_{k=1}^{m-1}\frac{1}{k}\right) \tag{5.6}$$

with $c = 25$.

Proof. Denote

$$\mathcal{L}(\epsilon, m, w) := \left\{\Lambda \subseteq [1,m] : \sum_{i\in\Lambda} w_i \ge 1/2 + \epsilon\right\}.$$

Then for any $\epsilon \in [0, 1/2]$, m, w we have $|\mathcal{L}(\epsilon, m, w))| \ge 1$. Therefore, (5.6) obviously holds for $m \le 6$, $\epsilon \in [0, 1/2]$ and for any $m > 6$, $\epsilon \in [\beta, 1/2]$, $\beta = (\ln 2)^{1/2}/5$.

We first establish Lemma 5.2 for $\epsilon \in [0, (9m)^{-1/2}]$. We will use a simple property of the Rademacher functions $\{r_i(t)\}$.

LEMMA 5.3. *Let* $\sum_{i=1}^n |c_i| = 1$. *Then*

$$\operatorname{mes}\Big\{t : \Big|\sum_{i=1}^n c_i r_i(t)\Big| \le 2(9n)^{-1/2}\Big\} \le 1 - 5/(9n).$$

Proof. Denote

$$g := \sum_{i=1}^n c_i r_i \quad \text{and} \quad E := \{t : |g(t)| \le 2(9n)^{-1/2}\}.$$

Then we have on the one hand

$$\|g\|_2^2 = \sum_{i=1}^n c_i^2 \ge 1/n. \tag{5.7}$$

On the other hand

$$\|g\|_2^2 \le (4/(9n))|E| + (1 - |E|). \tag{5.8}$$

Comparing (5.7) and (5.8) we get

$$|E| \le 1 - 5/(9n).$$

We continue the proof of Lemma 5.2 in the case $\epsilon \in [0, (9m)^{-1/2}]$. We observe that

$$\begin{aligned} 2^{-m}|\mathcal{L}(\epsilon, m, w)| &= \operatorname{mes}\Big\{t : \sum_{i=1}^m w_i(r_i(t) + 1)/2 \ge 1/2 + \epsilon\Big\} \\ &= \operatorname{mes}\Big\{t : \sum_{i=1}^m w_i r_i(t) \ge 2\epsilon\Big\}. \end{aligned} \tag{5.9}$$

Using Lemma 5.3 we obtain

$$2^{-m}|\mathcal{L}((9m)^{-1/2}, m, w)| \ge 5/(9m).$$

This inequality combined with the simple inequality

$$6\sum_{k=1}^{m-1} \frac{1}{k} \ge \ln(2m), \qquad m = 2, 3, \dots,$$

gives us (5.6) in the case $\epsilon \in [0, (9m)^{-1/2}]$.

It remains to consider the case $\epsilon \in [(9m)^{-1/2}, \beta]$. The proof of this case goes by induction. As we have already mentioned (5.6) holds for $m \le 6$. So, we assume that (5.6) holds for $m - 1$ and derive from it (5.6) for m. Denoting $w' := (w_1, \dots, w_{m-1})$, $w^1 := w'(1 - w_m)^{-1}$ we get

$$\mathcal{L}(\epsilon, m, w) = \{\{m\} \cup \Lambda, \Lambda \in \mathcal{L}(\epsilon - w_m, m - 1, w')\} \cup \mathcal{L}(\epsilon, m - 1, w'). \tag{5.10}$$

Next,

$$\begin{aligned} \mathcal{L}(\epsilon - w_m, m - 1, w') &= \mathcal{L}((\epsilon - w_m/2)(1 - w_m)^{-1}, m - 1, w^1), \\ \mathcal{L}(\epsilon, m - 1, w') &= \mathcal{L}((\epsilon + w_m/2)(1 - w_m)^{-1}, m - 1, w^1). \end{aligned}$$

Using the notations $\epsilon_1 := (\epsilon - w_m/2)(1 - w_m)^{-1}$, $\epsilon_2 := (\epsilon + w_m/2)(1 - w_m)^{-1}$ we obtain from (5.10)

$$|\mathcal{L}(\epsilon, m, w)| = |\mathcal{L}(\epsilon_1, m - 1, w^1)| + |\mathcal{L}(\epsilon_2, m - 1, w^1)|.$$

By the induction assumption we hence get

$$|\mathcal{L}(\epsilon,m,w)| \ge 2^{m-1}\exp\Big(-\frac{c}{4}\sum_{k=1}^{m-2}\frac{1}{k}\Big)\big(\exp(-c(m-1)\epsilon_1^2)+\exp(-c(m-1)\epsilon_2^2)\big).$$

We want to apply Lemma 5.1 with $n=m$. The assumptions of Lemma 5.1 $\epsilon\in(0,\beta]$, $m\ge(3\epsilon)^{-2}$ follow from $\epsilon\in[(9m)^{-1/2},\beta]$. Therefore, by Lemma 5.1 we obtain

$$|\mathcal{L}(\epsilon,m,w)| \ge 2^{m}\exp\Big(-cm\epsilon^2-\frac{c}{4}\sum_{k=1}^{m-1}\frac{1}{k}\Big).$$

This completes the proof of Lemma 5.2.

THEOREM 5.2. *For any $\epsilon\in[0,1/2]$, $m\ge 2$, and $w=(w_1,w_2,\dots,w_m)$ we have*

$$\Big|\Big\{\Lambda\subseteq[1,m]:\Big|\sum_{i\in\Lambda}w_i-1/2\Big|\ge\epsilon\Big\}\Big| \ge 2^m\exp\Big(-cm\epsilon^2-\frac{c}{4}\sum_{k=1}^{m-1}\frac{1}{k}\Big)$$

with $c=25$.

Proof. Denote

$$\mathcal{L}'(\epsilon,m,w):=\Big\{\Lambda\subseteq[1,m]:\Big|\sum_{i\in\Lambda}w_i-1/2\Big|\ge\epsilon\Big\}.$$

Similarly to (5.9) we have

$$2^{-m}|\mathcal{L}'(\epsilon,m,w)| = \operatorname{mes}\Big\{t:\Big|\sum_{i=1}^{m}w_i(r_i(t)+1)/2-1/2\Big|\ge\epsilon\Big\}. \tag{5.11}$$

Denoting $s:=\sum_{i=1}^m w_i$ we continue (5.11)

$$=\operatorname{mes}\Big\{t:\sum_{i=1}^{m}w_ir_i(t)\ge 1-s+2\epsilon\Big\}+\operatorname{mes}\Big\{t:\sum_{i=1}^{m}w_ir_i(t)\le 1-s-2\epsilon\Big\}$$

$$=\operatorname{mes}\Big\{t:\sum_{i=1}^{m}|w_i|r_i(t)\ge 1-s+2\epsilon\Big\}+\operatorname{mes}\Big\{t:\sum_{i=1}^{m}|w_i|r_i(t)\le 1-s-2\epsilon\Big\}=:M_1+M_2.$$

Denote $a:=\sum_{i=1}^m|w_i|$ and $u_i:=|w_i|/a$. In the case $a\ge 1$, $s\ge 1$ we have

$$M_1=\operatorname{mes}\Big\{t:\sum_{i=1}^{m}u_ir_i(t)\ge(1-s)/a+2\epsilon/a\Big\}\ge\operatorname{mes}\Big\{t:\sum_{i=1}^{m}u_ir_i(t)\ge 2\epsilon\Big\}.$$

We get the required estimate by Lemma 5.2. In the case $a\ge 1$, $s\le 1$ we get in the same way as above

$$M_2\ge\operatorname{mes}\Big\{t:\sum_{i=1}^{m}u_ir_i(t)\le -2\epsilon\Big\}. \tag{5.12}$$

By Lemma 5.2 we complete the case.

Let $0<a<1$. Then using $s\le a$ we get

$$(1-s)/a-2\epsilon/a\ge -2\epsilon$$

and, therefore, (5.12) holds also in this case. It remains to use Lemma 5.2.

Theorem 5.2 is now proved.

Theorem 5.1 is an immediate corollary of Theorem 5.2.

I am grateful to Professor Kwapień for the following remark concerning the proof of Theorem 5.1.

REMARK 5.1. One can use the paper [HK] in the proof of Theorem 5.1. This gives the estimate

$$\text{Prob}\Big\{\Big|\sum_{i=1}^{m} w_i y_i - 1/2\Big| \geq \epsilon\Big\} \geq \exp(-cm\epsilon^2 - 6 - \ln 8) \tag{5.13}$$

with $c = 128$.

Also, S. Kwapień has given an argument how to improve the constant c in (5.13) from 128 to 24.

References

[B] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation*, Z. Wahrscheinlichkeitstheorie Verw. Geb. 65 (1983), 181–237.

[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.

[BM1] L. Birgé and P. Massart, *Rates of convergence for minimum contrast estimators*, Probability Theory and Related Fields 97 (1993), 113–150.

[BM2] L. Birgé and P. Massart, *Minimax contrast estimators on sieves: exponential bounds and rates of convergence*, Bernoulli 4 (1998), 329–375.

[BM3] L. Birgé and P. Massart, *Gaussian model selection*, J. Eur. Math. Soc. 3 (2001), 203–268.

[BS] M. Sh. Birman and M. Z. Solomyak, *Piecewise polynomial approximation of the class* W_p^α, Mat. Sb. 73 (115) (1967), 331–355; English transl. in Math. USSR-Sb 2 (1967).

[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.

[D] R. DeVore, *Nonlinear approximation*, Acta Numer. 7 (1998), 51–150.

[DJ] D. Donoho and I. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika 81 (1994)), 425-455.

[DKPT1] R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov, *On mathematical methods of learning*, IMI Preprints 10 (2004), 1–24.

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

[G] U. Grenander, *Abstract Inference*, Wiley, New York, 1981.

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

[HK] P. Hitczenko and S. Kwapień, *On Rademacher series*, Progress in Probability 35 (1994), 31–36.

[IH] I. A. Ibragimov and R. Z. Hasminskii, *Statistical Estimation: Asymptotic Theory*, Springer, New York, 1981.

[KL] S. Kullback and R. A. Leibler, *On information and sufficiency*, Ann. Math. Statist. 22 (1951), 79–86.

[KP] G. Kerkyacharian and D. Picard, *Entropy, universal coding, approximation and bases properties*, Constructive Approximation 20 (2004), 1–37.

[KT1] S. Konyagin and V. Temlyakov, *Some error estimates in learning theory*, in: Approximation Theory: A volume dedicated to Borislav Bojanov, Marin Drinov Academic Publishing House, Sofia, 2004, 126–144.

[KT2] S. Konyagin and V. Temlyakov, *The entropy in the learning theory. Error estimates*, IMI Preprints 09 (2004), 1–25.

[L] G. Lugosi, *Pattern classification and learning theory*, in: Principles of Nonparametric Learning, Springer, Vienna, 2002, 5–62.

[M] S. Mendelson, *A few notes on statistical learning theory*, in: Advanced Lectures in Machine Learning, LNCS 2600, Springer, 2003, 1–40.

[P] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, 1989.

[Po] D. Pollard, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[PS] T. Poggio and S. Smale, *The mathematics of learning: dealing with data*, manuscript (2003), 1–16.

[S] C. J. Stone, *Optimal global rates of convergence for nonparametric regression*, Annals of Statistics 10 (1982), 1040–1053.

[T] V. N. Temlyakov, *Nonlinear methods of approximation*, Found. Comput. Math. 3 (2003), 33–107.

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

[VG] S. Van de Geer, *Empirical Process in M-Estimation*, Cambridge University Press, New York, 2000.

[YB] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics 27 (1999), 1564–1599.