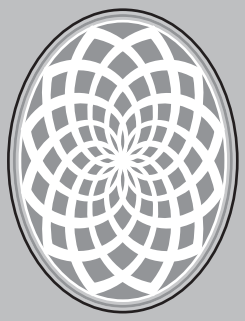

Industrial Mathematics Institute

2005:05

Approximation in learning theory

V. Temlyakov

IMI

Preprint Series

Department of Mathematics
University of South Carolina

APPROXIMATION IN LEARNING THEORY

V.N. Temlyakov

Abstract. This paper addresses 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 (or its projection) 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 the following two problems: I. the projection learning problem (improper function learning problem); II. universal (adaptive) estimators in the proper function learning problem. In the first problem we do not impose any restrictions on a Borel measure ρ except our standard assumption that $|y| \le M$ a.e. with respect to ρ. In this case we use the data $\mathbf{z}$ to estimate (approximate) the $L_2(\rho_X)$ projection $(f_\rho)_W$ of f_ρ onto a function class W of our choice. Here, ρ_X is the marginal prabability measure. In [KT1,2] this problem has been studied for W satisfying the decay condition $\epsilon_n(W, B) \le Dn^{-r}$ of the entropy numbers $\epsilon_n(W, B)$ of W in a Banach space B in the case $B = \mathcal{C}(X)$ or $B = L_2(\rho_X)$. In this paper we obtain the upper estimates in the case $\epsilon_n(W, L_1(\rho_X)) \le Dn^{-r}$ with an extra assumption that W is convex.

In the second problem we assume that an unknown measure ρ satisfies some conditions. Following the standard way from nonparametric statistics we formulate these conditions in the form $f_\rho \in \Theta$. Next, we assume that the only a priori information available is that f_ρ belongs to a class Θ (unknown) from a known collection $\{\Theta\}$ of classes. We want to build an estimator that provides approximation of f_ρ close to the optimal for the class Θ. Along with standard penalized least squares estimators we consider a new method of construction of universal estimators. This method is based on a combination of two powerful ideas in building universal estimators. The first one is the use of penalized least squares estimators. This idea works well in the case of general setting with rather abstract methods of approximation. The second one is the idea of thresholding that works very well when we use wavelets expansions as an approximation tool. A new estimator that we call *big jump estimator* uses the least squares estimators and chooses a right model by a thresholding criteria instead of the penalization. In this paper we illustrate how ideas and methods of approximation theory can be used in learning theory both in formulation of a problem and in solving it.

1. Introduction. Setting. Known results

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. This is a big area of research both in nonparametric statistics and in learning theory. In this paper we confine ourselves to results obtained in a direction of further development of the settings and results from the paper of Cucker and Smale [CS]. For results in other settings we recommend a book of V. Vapnik [V], a survey by T. Evgeniou, M. Pontil and T. Poggio [EPP], and a survey on the classification problem by G. Lugosi [L]. Our setting is similar to the setting of the distribution-free regression problem. In this paper we illustrate how ideas and methods of approximation theory can be used in learning theory both in formulation of a problem and in solving it.

A standard mathematical framework for the setting of the above learning problem is the following ([CS], [PS], [DKPT1,2],[KT1,2], [T2]). Let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ be Borel sets, ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \to Y$ define *the error*

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

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

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

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 $f_{\mathbf{z}}$, 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 accoding to ρ.

There are several important ingredients in mathematical formulation of this problem. We follow the way that has become standard in approximation theory and has been used in [DKPT1,2], [KT1,2], and [T2]. In this approach we first 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], [KT1,2], [T2]) is based on the idea of studying approximation of the $L_2(\rho_X)$ projection $f_W := (f_\rho)_W$ of f_ρ onto W. This setting is known as the *improper function learning problem* or the *projection learning problem*. 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], [DKPT1,2], [KT1,2], [BCDDT], [T2]) is based on the assumption $f_\rho \in W$. This setting is known as the *proper function learning problem*. For instance, we may assume that f_ρ has some smoothness. In the case of the proper function learning problem we use the notation Θ (instead of W) for a class of priors. In Sections 2 and 3 of this paper we study the projection learning problem and in Sections 4–6 we study the proper function learning problem.

The main question of nonparametric regression theory and learning theory is how to choose an estimator $f_{\mathbf{z}}$. There are several different approaches to this problem. We now

discuss some of them. Recently, driven by ideas from approximation theory, the following general approach to this problem has been developed. The idea of this approach is to choose an estimator $f_{\mathbf{z}}$ as a solution (approximate solution) of an optimization problem (minimax problem). So, in this approach we should begin with a formulation of an optimization problem. A standard formulation of such a problem is the following. We begin with a fixed class Θ of priors (or a fixed class W where we project f_ρ). That means we impose a restriction on an unknown measure ρ, which we want to study, in the form $f_\rho \in \Theta$. Developping this approach we encounter three immediate questions: 1. What classes Θ of priors (or classes W) to choose? 2. What should be the form of $f_{\mathbf{z}}$? 3. How to measure the quality of estimation (approximation)? We will not discuss these questions in detail here. We only note that the following partial answers to the above questions are widely accepted. 1. A very important characteristic of Θ that governs the quality of estimation is a sequence of the entropy numbers $\epsilon_n(\Theta, B)$ of Θ in a suitable Banach space B. 2. The following way of building $f_{\mathbf{z}}$ provides a near optimal estimator in many cases. First, choose a right hypothesis space $\mathcal{H}$ (that may depend on Θ). Second, construct $f_{\mathbf{z},\mathcal{H}} \in \mathcal{H}$ as the empirical optimum (least squares estimator). We explain this in more detail. 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 closed unit ball of a Banach space B. We denote $N(\Theta, \epsilon, B)$ the covering number that is the minimal number of balls of radius ϵ with centers in Θ needed for covering Θ. We note that $N(\Theta, \epsilon_n(\Theta, B), B) \le 2^n$.

We define

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

where

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

is the *empirical error (risk)* of f. This $f_{\mathbf{z},\mathcal{H}}$ is called the *empirical optimum* or the *least squares estimator*. 3. It seems natural (see [CS], [GKKW], [DKPT1], [DKPT2], [KT1], [KT2], [BCDDT]) to measure the quality of approximation by $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$. It is easy to see that for any $f \in L_2(\rho_X)$

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

Thus the choice $\|\cdot\| = \|\cdot\|_{L_2(\rho_X)}$ looks natural. The reader can find a discussion of recent results on optimal rates of estimation for different classes Θ in a survey [T2].

In this paper we address the following important issue. In many cases we do not know exactly what is a class Θ of priors where an unknown f_ρ comes from. Therefore, we try to construct an estimator that provides good estimation (near optimal) not for a single class of priors Θ but for a collection of classes of priors. Clearly, in order to claim that an estimator $f_{\mathbf{z}}$ is near optimal for a class Θ we need to compare the upper estimates of

approximation by $f_{\mathbf{z}}$ with the corresponding lower bounds of optimal estimation for Θ. We will discuss some known lower estimates. The usual in regression theory 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 $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)}$. A more accurate and more delicate way of evaluating the performance of $f_{\mathbf{z}}$ has been pushed forward in [CS]. In [CS] they study the probability distribution function

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

instead of the expectation $E(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)})$.

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

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

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

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

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

We begin our discussion of known results with the lower estimate of the accuracy confidence function from [DKPT2]. These lower bounds have been established in terms of a certain variant of the Kolmogorov entropy of Θ which we shall call *tight entropy*. For a compact Θ in a Banach space B we define the *packing numbers* as

$$P(\Theta, \delta) := P(\Theta, \delta, B) := \sup\{N : \exists\ f_1, ..., f_N \in \Theta,$$

with

$$\delta \le \|f_i - f_j\|_B,\ \forall i \ne j\}.$$

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

$$\bar{P}(\Theta, \delta) := \bar{P}(\Theta, \delta, c_1, B) := \sup\{N : \exists\ f_1, ..., f_N \in \Theta,$$

with

$$\delta \le \|f_i - f_j\|_B \le c_1\delta,\ \forall i \ne j\}. \tag{1.1}$$

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

We let μ be any Borel probability measure defined on X and let $\mathcal{M}(\Theta, \mu)$ denote the set of all ρ such that $\rho_X = \mu$, $|y| \le 1$, $f_\rho \in \Theta$. 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 probability 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 (1.1). 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 $\mathcal{M}(\Theta, \mu)$.

The following theorems are known.

Theorem 1.1 [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 (1.1) 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 $i \in \{1, \dots, \bar{P}\}$*

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

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 $b_n \ll a_n$.

Theorem 1.2 [T2]. *Assume Θ is a compact subset of $L_2(\mu)$ such that $\Theta \subset \frac{1}{4}U(\mathcal{C}(X))$ and*

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}. \tag{1.2}$$

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

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \ge \delta_0 \quad \textit{for} \quad \eta \le \eta_m \tag{1.3}$$

and

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \ge Ce^{-c(r)m\eta^2} \quad \textit{for} \quad \eta \ge \eta_m. \tag{1.4}$$

Remark 1.1 [T2]. *Theorem 1.2 holds in the case* $\Theta \subset \frac{M}{4}U(\mathcal{C}(X))$, $|y| \le M$, *with constants allowed to depend on* M.

The lower estimates from Theorem 1.2 will serve as a benchmark for the performance of particular estimators. Let us formulate a condition on a measure ρ and a class $\mathcal{H}$ that we will often use:

$$\text{for all} \quad f \in \mathcal{H}, \quad \text{we have} \quad |f(x) - y| \le M \quad \text{a.e. with respect to} \quad \rho. \tag{1.5}$$

Clearly, (1.5) is satisfied if $|y| \le M/2$ and $|f(x)| \le M/2$, $f \in \mathcal{H}$.

In Section 4 we prove the following complementary to Theorem 1.2 result.

Theorem 1.3. *Let* $f_\rho \in \Theta$ *and let* ρ, Θ *satisfy (1.5). Assume*

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

Then there exists an estimator $f_{\mathbf{z}}$ *such that for* $\eta \ge 7\epsilon_0$, $\epsilon_0 := C(M, D, r)m^{-\frac{2r}{1+2r}}$, $m \ge 60(M/D)^2$, *we have*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)} \ge \eta\} \le \exp\Big(-\frac{m\eta}{200M^2}\Big).$$

In the case of Θ satisfying the assumption

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

Theorem 1.3 has been obtained in [KT1].

A combination of Theorems 1.2 and 1.3 completes the study of the behavior (in the sense of order) of the **AC**-function of classes satisfying (1.2). We formulate this as a theorem.

Theorem 1.4. *Let* μ *be a Borel probability measure on* X. *Assume* $r > 0$ *and* Θ *is a compact subset of* $L_2(\mu)$ *such that*

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

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

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

and

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

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

Let us make a general remark on the technique that we use in this paper. It usually consists of a combination of results from nonparametric statistics with results from approximation theory. Both the results from nonparametric statistics and the results from

approximation theory that we use are either known or very close to known results. The novelty of this paper is in combining these known results and applying them in a new setting. For example, in the proof of Theorem 1.3 we have used a statistical technique that was used in many papers (for instance, [LBW], [BBM], [CS]) and goes back to Barron's seminal paper [B]. We also used some elementary results on the entropy numbers from approximation theory.

We now proceed to results on construction of universal (adaptive) estimators. Let us begin with a case where we impose conditions on the class Θ in a spirit of Kolmogorov's widths. Denote for a set L of a Banach space B

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

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

$$d(W^r(\mathcal{L}), L_n)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n=1,2,\dots; \quad W^r(\mathcal{L}) \subset DU(\mathcal{C}(X)).$$

In the following discussion let us 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) 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 the introduction without formulating it. In Sections 4 and 7 we prove the following theorem (see Theorem 7.1) that extends the corresponding results from [DKPT1,2] for the collection $\mathcal{W}(\mathcal{L},\alpha,1/2)$ to a result for the collection $\mathcal{W}(\mathcal{L},\alpha,\beta)$ with any $0<\alpha\le\beta<\infty$.

Theorem 1.5. *For a given collection $\mathcal{W}(\mathcal{L},\alpha,\beta)$, $0<\alpha\le\beta<\infty$, there exists an estimator $f_{\mathbf{z}}$ such that if $f_\rho\in W^r(\mathcal{L})$, $r\in[\alpha,\beta]$ then for $A\ge A_0(M,\alpha,\beta)$*

$$\rho^m\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)\le C(D)A(\ln m/m)^{\frac{2r}{1+2r}}\} \ge 1-m^{C_1(M)(C_2(M)-A)}.$$

The above theorem provides a universal estimator for the collection $\mathcal{W}(\mathcal{L},\alpha,\beta)$ of classes defined in terms of approximation in the uniform norm by linear subspaces. As we already mentioned the natural norm to work with in learning theory is the $L_2(\rho_X)$ norm.

We do not know if Theorem 1.5 holds for f_ρ satisfying

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

instead of $f_\rho\in W^r(\mathcal{L})$. However, we prove in Section 4 such a generalization of Theorem 1.5 with the subspaces L_n, $n=1,2,\dots$ replaced, for instance, by the $\mathcal{C}(X)$-balls of L_n. We note that these results generalize the corresponding results from [KT2] where we imposed extra assumptions on the subspaces L_n in the form of uniform boundedness of the $L_2(\mu)$ projections on L_n as operators from $\mathcal{B}(X)$ to $\mathcal{B}(X)$ (see Section 7 for detail). We also mention the paper [BCDDT] that addresses in addition to the issue of universality the issue

of online implementation. The construction of good estimators from [BCDDT] is based on adaptive (data dependent) partitioning of X and on thresholding.

In this paper we give a general scheme for construction of universal estimators. It begins with a sequence of hypothesis spaces $\mathcal{H}_n$, $n = 1, 2, \dots$. Then we consider the sequence of least squares estimators $f_{\mathbf{z},\mathcal{H}_n}$, $n = 1, 2, \dots$. Next, we use two different ways of choosing an estimator $f_{\mathbf{z}} := f_{\mathbf{z},\mathcal{H}_{n(\mathbf{z})}}$ (choosing $n(\mathbf{z})$). The first way is based on a known idea of penalization. We discuss the corresponding results in Section 4. The second way is based on a thresholding type criterion for the differences $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_{2^{s-1}}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_{2^s}})$. We discuss this way in Sections 5 and 6.

In Sections 5 and 6 we consider a new method of construction of universal estimators. This method is based on a combination of two powerful ideas in building universal estimators. The first one is the use of penalized least squares estimators. This idea works well in the case of general setting with rather abstract methods of approximation, like in Section 4. The second one is the idea of thresholding that works very well when we use wavelets expansions as an approximation tool. A new estimator that we call *big jump estimator* uses the least squares estimators and chooses a right model by thresholding criteria instead of penalization. The technique of studying these new estimators is more complicated than the technique developed in Section 4 for studying the penalized least aquares estimators. As a result we got some restrictions, for instance, $r \le 1/2$ in Theorem 5.2, that probably reflect technical difficulties rather than a new phenomenon. Our method uses a mixture of the previous techniques: we measure compactness of $\mathcal{H}_n$ in the uniform norm and approximate f by elements from $\mathcal{H}_n$ in the $L_2(\rho_X)$ norm.

Section 3 plays a preparatory role for Sections 5–6. However, it might be of independent interest. In this section we generalize a known result from [CS] that holds for convex hypothesis spaces $\mathcal{H}$ to the case of nonconvex hypothesis spaces $\mathcal{H}$. We prove that the condition of convexity can be replaced by a control of the distance from f_ρ to $\mathcal{H}$.

We will often use the classical Bernstein's inequalities. If ξ is a random variable (a real valued function on a probability space Z) then denote

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

For a single function f we have the following probabilistic Bernstein inequalities. If $|\xi(z) - E(\xi)| \le M$ a.e. then for any $\epsilon > 0$ one has

$$\text{Prob}_{\mathbf{z}\in Z^m}\{|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)| \ge \epsilon\} \le 2\exp\left(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}\right) \tag{1.6}$$

with $\mathbf{z} := (z_1, \dots, z_m)$. Here are the corresponding one-sided inequalities

$$\text{Prob}_{\mathbf{z}\in Z^m}\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi) \ge \epsilon\} \le \exp\left(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}\right), \tag{1.7}$$

$$\text{(1.8)}\qquad \text{Prob}_{\mathbf{z}\in Z^m}\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi) \le -\epsilon\} \le \exp\left(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}\right).$$

In the paper we formulate assumptions on a class W in the following form

$$\text{(1.9)}\qquad \epsilon_n(W, B) \le Dn^{-r}, \quad W \subset DU(B).$$

We denote $\mathcal{S}^r := \mathcal{S}^r(D) := \mathcal{S}^r_\infty(D)$ a collection of classes W that satisfy (1.9) with $B = \mathcal{C}(X)$. The notation $\mathcal{S}^r_p := \mathcal{S}^r_p(D)$ is used for a collection of classes W satisfying (1.9) with $B = L_p(\rho_X)$, $1 \le p < \infty$.

We often have error estimates of the form $(\ln m/m)^\alpha$ that hold for $m \ge 2$. 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. Least squares estimators for convex hypothesis spaces

In this section we give some results from [CS] in the case when the hypothesis space $\mathcal{H}$ is convex. We present these results with complete proofs for the following reason. In Section 3 we proof similar results under the assumption of convexity replaced by other assumption. In Section 4 we formulate results similar to those from this section without proofs. So, for the sake of completeness and for the sake of the reader's convenience we have included the complete proofs here. Also, our proofs that use the same ideas as in [CS] are a little simpler technically and give better numerical constants in the inequalities. We note that the technique presented in this section is a development of techniques used in [B], [LBW], [BBM]. At the end of Section 2 we apply this technique to a problem of the projection learning (improper function learning). Theorem 2.5 is new, it extends known results (Theorem 2.4) in the direction of replacing the assumption $\mathcal{H} \in \mathcal{S}^r$ by a weaker assumption $\mathcal{H} \in \mathcal{S}^r_1$. Further comments are given at the end of Section 2. In addition to the notation $f_{\mathbf{z},\mathcal{H}}$ defined in the Introduction we will use the following notation for the $L_2(\rho_X)$ projection of f_ρ onto $\mathcal{H}$ (we assume existence)

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

We begin with the following theorem.

Theorem 2.1 [CS]. *Suppose that $\mathcal{H}$ is a compact and convex subset of $\mathcal{C}(X)$. Assume that ρ and $\mathcal{H}$ satisfy (1.5). Then, for all $\epsilon > 0$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) < \epsilon\} \ge 1 - N(\mathcal{H}, \epsilon/(16M))\exp(-\frac{m\epsilon}{80M^2}).$$

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

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

Proof. By the convexity assumption for any $f \in \mathcal{H}$ and $g := f - f_{\mathcal{H}}$, we have $(1-\epsilon)f_{\mathcal{H}} + \epsilon f = f_{\mathcal{H}} + \epsilon g$ is in $\mathcal{H}$ and therefore,

$$0 \le \|f_\rho - f_{\mathcal{H}} - \epsilon g\|^2_{L_2(\rho_X)} - \|f_\rho - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} = -2\epsilon \int_X (f_\rho - f_{\mathcal{H}}) g \, d\rho_X + \epsilon^2 \int_X g^2 \, d\rho_X.$$

Letting $\epsilon \to 0$, we obtain the following well-known result:

$$\int_X (f_\rho - f_{\mathcal{H}})(f - f_{\mathcal{H}}) \, d\rho_X \le 0, \quad f \in \mathcal{H}. \tag{2.1}$$

Then letting $\epsilon = 1$ we see that $\|f_\rho - f\|_{L_2(\rho_X)} > \|f_\rho - f_{\mathcal{H}}\|_{L_2(\rho_X)}$ whenever $f \ne f_{\mathcal{H}}$ and so $f_{\mathcal{H}}$ is unique. Also, (2.1) gives

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

We will use the following notations.

$$\delta(\mathcal{H}) := \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) = \|f_{\mathcal{H}} - f_\rho\|^2_{L_2(\rho_X)};$$

$$\mathcal{E}_{\mathcal{H}}(f) := \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}); \qquad \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}});$$

$$\ell(f) := \ell(f,z) := (f(x) - y)^2 - (f_{\mathcal{H}}(x) - y)^2, \quad z = (x,y).$$

We note that

$$\mathcal{E}_{\mathcal{H}}(f) = E_\rho(\ell(f,z)); \qquad \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m \ell(f,z_i).$$

Lemma 2.2 [CS]. *Assume that $\mathcal{H}$ is convex and ρ and $\mathcal{H}$ satisfy (1.5). Then we have*

$$\sigma^2 := \sigma^2(\ell(f)) \le 4M^2 \mathcal{E}_{\mathcal{H}}(f).$$

Proof. We have

$$\sigma^2(\ell(f)) \le E(\ell(f)^2) = E((f(x) - f_{\mathcal{H}}(x))^2 (f(x) + f_{\mathcal{H}}(x) - 2y)^2)$$

$$\le 4M^2 E((f(x) - f_{\mathcal{H}}(x))^2) = 4M^2 \|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le 4M^2 \mathcal{E}_{\mathcal{H}}(f).$$

At the last step we have used Lemma 2.1. □

Lemma 2.3 [CS]. *Assume that $\mathcal{H}$ is convex and ρ and $\mathcal{H}$ satisfy (1.5). Let $f \in \mathcal{H}$. For all $\epsilon > 0$, $\alpha \in (0,1]$ one has*

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \geq \alpha(\mathcal{E}_{\mathcal{H}}(f) + \epsilon)\} \leq \exp\big(-\frac{\alpha^2 m\epsilon}{5M^2}\big),$$

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \leq -\alpha(\mathcal{E}_{\mathcal{H}}(f) + \epsilon)\} \leq \exp\big(-\frac{\alpha^2 m\epsilon}{5M^2}\big).$$

Proof. Denote $a := \mathcal{E}_{\mathcal{H}}(f)$. We note that $a \geq 0$. The proofs of both inequalities are the same. We will carry out only the proof of the first one. Using one-sided Bernstein's inequality (1.8) for $\ell(f)$ we obtain

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \geq \alpha(a+\epsilon)\} \leq \exp\big(-\frac{m\alpha^2(a+\epsilon)^2}{2(\sigma^2 + M^2\alpha(a+\epsilon)/3)}\big).$$

It remains to check that

$$\frac{(a+\epsilon)^2}{2(\sigma^2 + M^2\alpha(a+\epsilon)/3)} \geq \frac{\epsilon}{5M^2}. \tag{2.1}$$

Using Lemma 2.2 we get on the one hand

$$2\epsilon(\sigma^2 + M^2\alpha(a+\epsilon)/3) \leq M^2\epsilon(9a + 2\epsilon/3). \tag{2.2}$$

On the other hand

$$5M^2(a+\epsilon)^2 \geq M^2\epsilon(10a + 5\epsilon). \tag{2.3}$$

Comparing (2.2) and (2.3) we obtain the required inequality. □

Lemma 2.4 [CS]. *Assume that ρ and $\mathcal{H}$ satisfy (1.5). Let $\alpha \in (0,1)$, $\epsilon > 0$, and let $f \in \mathcal{H}$ be such that*

$$\frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f) + \epsilon} < \alpha. \tag{2.4}$$

Then for all $g \in \mathcal{H}$ such that $\|f - g\|_{C(X)} \leq \frac{\alpha\epsilon}{4M}$ we have

$$\frac{\mathcal{E}_{\mathcal{H}}(g) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathcal{H}}(g) + \epsilon} < 2\alpha. \tag{2.5}$$

Proof. Denote

$$a := \mathcal{E}_{\mathcal{H}}(f), \quad a' := \mathcal{E}_{\mathcal{H}}(g), \quad b := \mathcal{E}_{\mathcal{H},\mathbf{z}}(f), \quad b' := \mathcal{E}_{\mathcal{H},\mathbf{z}}(g).$$

Then our assumption $\|f-g\|_{\mathcal{C}(X)} \le \frac{\alpha\epsilon}{4M}$ implies

$$|a-a'| \le \alpha\epsilon/2, \qquad |b-b'| \le \alpha\epsilon/2. \tag{2.6}$$

By (2.4) and (2.6) we get ($a \ge 0$)

$$a(1-\alpha) < b + \alpha\epsilon \le b' + 3\alpha\epsilon/2. \tag{2.7}$$

Also, by (2.6)

$$a(1-\alpha) \ge (a' - \alpha\epsilon/2)(1-\alpha) \ge a' - \alpha a' - \alpha\epsilon/2. \tag{2.8}$$

Combining (2.7) and (2.8) we obtain

$$a' - b' < \alpha a' + 2\alpha\epsilon \le 2\alpha(a' + \epsilon)$$

which implies (2.5). □

A combination of Lemma 2.3 and Lemma 2.4 gives the following theorem.

Theorem 2.2 [CS]. *Assume that $\mathcal{H}$ is convex and ρ, $\mathcal{H}$ satisfy (1.5). Then for all $\epsilon > 0$ and $\alpha \in (0,1)$*

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f)+\epsilon} \ge 2\alpha\} \le N(\mathcal{H}, \frac{\alpha\epsilon}{4M}, \mathcal{C}(X)) \exp\big(-\frac{\alpha^2 m\epsilon}{5M^2}\big).$$

Proof. Let $f_1, \dots, f_N$ be the $\frac{\alpha\epsilon}{4M}$-net of $\mathcal{H}$ in $\mathcal{C}(X)$, $N := N(\mathcal{H}, \frac{\alpha\epsilon}{4M}, \mathcal{C}(X))$. Let Λ be the set of $\mathbf{z}$ such that for all $j = 1, \dots, N$ we have

$$\frac{\mathcal{E}_{\mathcal{H}}(f_j) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f_j)}{\mathcal{E}_{\mathcal{H}}(f_j)+\epsilon} < \alpha.$$

Then by Lemma 2.3

$$\rho^m(\Lambda) \ge 1 - N\exp\big(-\frac{\alpha^2 m\epsilon}{5M^2}\big). \tag{2.9}$$

We take any $\mathbf{z} \in \Lambda$ and any $g \in \mathcal{H}$. Let f_j be such that $\|g - f_j\|_{\mathcal{C}(X)} \le \frac{\alpha\epsilon}{4M}$. By Lemma 2.4 we obtain that

$$\frac{\mathcal{E}_{\mathcal{H}}(g) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathcal{H}}(g)+\epsilon} < 2\alpha.$$

It remains to use (2.9). □

Theorem 2.3 [CS]. *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$ and ρ, $\mathcal{H}$ satisfy (1.5). Then for all $\epsilon > 0$ with probability at least*

$$p(\mathcal{H}, \epsilon) := 1 - N(\mathcal{H}, \frac{\epsilon}{16M}, \mathcal{C}(X)) \exp\big(-\frac{m\epsilon}{80M^2}\big)$$

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

$$\mathcal{E}(f) \le 2\mathcal{E}_{\mathbf{z}}(f) + \epsilon - \mathcal{E}(f_{\mathcal{H}}) + 2(\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})). \tag{2.10}$$

Proof. Using Theorem 2.2 with $\alpha = 1/4$ we get with probability at least $p(\mathcal{H}, \epsilon)$

$$\mathcal{E}_{\mathcal{H}}(f) < 2\mathcal{E}_{\mathcal{H},\mathbf{z}}(f) + \epsilon. \tag{2.11}$$

Substituting

$$\mathcal{E}_{\mathcal{H}}(f) := \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}); \qquad \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})$$

we obtain (2.10). □

Corollary 2.1. *Under the assumptions of Theorem 2.3 we have*

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}}) \le \epsilon/2$$

with probability at least $p(\mathcal{H}, \epsilon)$.

We will use Corollary 2.1 and other corollaries of Theorem 2.3 in Section 5. The reader can find a proof of Corollary 2.1 in Section 5.

Proof of Theorem 2.1. We use (2.11) with $f = f_{\mathbf{z},\mathcal{H}}$. From the definition of $f_{\mathbf{z},\mathcal{H}}$ we obtain that $\mathcal{E}_{\mathcal{H},\mathbf{z}}(f_{\mathbf{z},\mathcal{H}}) \le 0$. This completes the proof of Theorem 2.1. □

The following theorem is a direct corollary of Theorem 2.1.

Theorem 2.4 [CS], [DKPT1,2]. *Assume that $\mathcal{H} \in \mathcal{S}^r$ is convex and ρ, $\mathcal{H}$ satisfy (1.5). Then for $\eta \ge \eta_m := A_0(M, D, r)m^{-\frac{r}{1+r}}$ one has*

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

Proof. We get from the assumption $\mathcal{H} \in \mathcal{S}^r$ that

$$N(\mathcal{H}, \eta/(16M)) \le 2^{(CDM/\eta)^{1/r}}.$$

Expressing η in the form $\eta = Am^{-\frac{r}{1+r}}$ we obtain that for $A \ge A_0(M, D, r)$

$$2^{(CDM/\eta)^{1/r}} \exp(-\frac{m\eta}{80M}) \le \exp(-c(M)m\eta).$$

Theorem 2.5. *Let W be a convex and compact in $L_1(\rho_X)$ set and let ρ, W satisfy (1.5). Assume $W \in \mathcal{S}_1^r$ that is*

$$\epsilon_n(W, L_1(\rho_X)) \le Dn^{-r}, \quad n = 1, 2, \dots, \quad W \subset DU(L_1(\rho_X)). \tag{2.12}$$

Then there exists an estimator $f_{\mathbf{z}} \in W$ such that for $\eta \ge \eta_m := (6M+4)\epsilon_0$, $\epsilon_0 := C(M, D, r)m^{-\frac{r}{1+r}}$, $m \ge 60(M/D)^2$, we have

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

Proof. Let $\mathcal{N} := \mathcal{N}_{\epsilon_0}(W, L_1(\rho_X))$ be a minimal ϵ_0-net of W in the $L_1(\rho_X)$ norm. The constant $C(M, D, r)$ will be chosen later. Then (2.12) implies that

$$|\mathcal{N}| \le 2^{(D/\epsilon_0)^{1/r}+1}. \tag{2.13}$$

As an estimator $f_{\mathbf{z}}$ we take

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

We take $\epsilon \ge \epsilon_0$ and apply the first inequality of Lemma 2.3 with $\alpha = 1/2$ to each $f \in \mathcal{N}$. In such a way we obtain a set Λ_1 with

$$\rho^m(\Lambda_1) \ge 1 - |\mathcal{N}| \exp\bigl(-\frac{m\epsilon}{20M^2}\bigr)$$

with the property: for all $f \in \mathcal{N}$ and all $\mathbf{z} \in \Lambda_1$ one has

$$\mathcal{E}_W(f) \le 2\mathcal{E}_{W,\mathbf{z}}(f) + \epsilon. \tag{2.14}$$

Therefore, for $\mathbf{z} \in \Lambda_1$

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathbf{z}}) + \epsilon \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) + \epsilon. \tag{2.15}$$

Let Λ_2 be the set of all $\mathbf{z}$ such that

$$\mathcal{E}_W(f_{\mathcal{N}}) - \mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) \le -\frac{1}{2}(\mathcal{E}_W(f_{\mathcal{N}}) + \epsilon). \tag{2.16}$$

By the second inequality of Lemma 2.3 with $\alpha = 1/2$

$$\rho^m(\Lambda_2) \le \exp\bigl(-\frac{m\epsilon}{20M^2}\bigr).$$

Consider $\Lambda := \Lambda_1 \setminus \Lambda_2$. Then

$$\rho^m(\Lambda) \ge 1 - (|\mathcal{N}| + 1)\exp\bigl(-\frac{m\epsilon}{20M^2}\bigr).$$

Using the inequality opposite to (2.16) we continue (2.15) for $\mathbf{z} \in \Lambda$

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) + \epsilon \le 3\mathcal{E}_W(f_{\mathcal{N}}) + 2\epsilon.$$

Next,

$$\mathcal{E}_W(f_{\mathcal{N}}) = \mathcal{E}(f_{\mathcal{N}}) - \mathcal{E}(f_W) = \min_{f\in\mathcal{N}}(\mathcal{E}(f) - \mathcal{E}(f_W)) \le \min_{f\in\mathcal{N}} 2M\|f - f_W\|_{L_1(\rho_X)} \le 2M\epsilon_0. \tag{2.17}$$

Therefore

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 6M\epsilon_0 + 2\epsilon.$$

We choose $\epsilon_0 \le D$ from the equation

$$3(D/\epsilon_0)^{\frac{1}{r}} = \frac{m\epsilon_0}{20M^2}.$$

We get

$$\epsilon_0 = (60M^2)^{\frac{r}{1+r}} D^{\frac{1}{1+r}} m^{-\frac{r}{1+r}}.$$

For $m \ge 60M^2/D$ we have $\epsilon_0 \le D$. We let $\eta = 6M\epsilon_0 + 2\epsilon$. Then our assumption $\eta \ge (6M+4)\epsilon_0$ implies $\epsilon \ge 2\epsilon_0$ and

$$\rho^m(Z^m \setminus \Lambda) \le (|\mathcal{N}| + 1)\exp\big(-\frac{m\epsilon_0}{20M^2}\big)\exp\big(-\frac{m(\epsilon - \epsilon_0)}{20M^2}\big)$$

$$\le \exp\big(-\frac{m\epsilon}{40M^2}\big) \le \exp\big(-\frac{m\eta}{40M^2(3M+2)}\big).$$

This completes the proof of Theorem 2.5. □

We note that Theorem 2.5 with the assumption $W \in \mathcal{S}^r$ (instead of $W \in \mathcal{S}_1^r$) has been proved in [CS], [DKPT1,2] with $f_{\mathbf{z}} = f_{\mathbf{z},W}$ (see Theorem 2.4). It is interesting to compare Theorem 2.5 with the corresponding results when we do not assume that W is convex. Let us compare only the accuracy thresholds η_m. Theorem 2.5 says that for a convex W the assumption $W \in \mathcal{S}_1^r$ implies

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

The results of [KT2] state that $W \in \mathcal{S}_2^r$ (no convexity assumption) implies

$$\eta_m \ll m^{-\frac{r}{1+r}}, \quad r \in (0,1],$$

$$\eta_m \ll m^{-1/2}, \quad r \ge 1.$$

The results of [KT1] give the following estimates for $W \in \mathcal{S}^r$

$$\eta_m \ll m^{-r}, \quad r \in (0, 1/2], \tag{2.18}$$

$$\eta_m \ll m^{-1/2}, \quad r \ge 1/2. \tag{2.19}$$

It has been proved in [KT1] that the estimates (2.18) and (2.19) cannot be improved. Therefore, even under a strong assumption $W \in \mathcal{S}^r$ the best we can get is $\eta_m \ll m^{-1/2}$. Theorem 2.5 shows that the convexity combined with a weaker assumption $W \in \mathcal{S}^r_1$ provide better estimates for big r.

Let us make a comment on studying the accuracy confidence function for the projection learning problem (improper function learning problem). Similarly to the case of the proper function learning problem we introduce the corresponding accuracy confidence function

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

where $\sup_\rho$ is taken over ρ such that ρ, W satisfy (1.5). 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 by Lemma 2.1 for any $f \in W$

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

Theorem 2.5 provides an upper estimate for the $\mathbf{AC}^p$-function in the case of convex W from $\mathcal{S}^r_1$:

$$\mathbf{AC}^p_m(W,\eta^{1/2}) \le \exp(-c(M)m\eta), \quad \eta \ge \eta_m \gg m^{-\frac{r}{1+r}}.$$

We note that the behavior of the $\mathbf{AC}^p$-function is well understood only in the following special cases. Let $r > 1/2$ then (see [KT1], [T2])

$$C_1\exp(-c_1(M)m\eta^4) \le \sup_{W\in\mathcal{S}^r(D)} \mathbf{AC}^p_m(W,\eta) \le C(M,D,r)\exp(-c_2(M)m\eta^4)$$

for $\eta \ge m^{-1/4}$. Also for $r \ge 1$ (see [KT2])

$$C_1\exp(-c_1(M)m\eta^4) \le \sup_{W\in\mathcal{S}^r_2(D)} \mathbf{AC}^p_m(W,\eta) \le C(M,D,r)\exp(-c_3(M)m\eta^4)$$

provided $\eta \gg m^{-1/4}$.

It would be interesting to find the behavior of

$$\sup_{W\in\mathcal{S}} \mathbf{AC}^p_m(W,\eta)$$

in the following cases: I. $\mathcal{S} = \mathcal{S}^r(D)$, $r \le 1/2$; II. $\mathcal{S} = \mathcal{S}^r_2(D)$, $r < 1$; III. $\mathcal{S} = \{W : W \in \mathcal{S}^r_q(D), \quad W \quad \text{is convex}\}$, $q = 1, 2, \infty$.

3. Least squares estimators for nonconvex hypothesis spaces

The following result has been proved in [DKPT1].

Theorem 3.1 [DKPT1]. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that ρ and $\mathcal{H}$ satisfy (1.5). Then, for all $\epsilon > 0$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \le \epsilon\} \ge 1 - N(\mathcal{H}, \epsilon/(24M))2\exp(-\frac{m\epsilon}{C(M,K)})$$

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

Theorem 3.1 shows that we obtain an analogue of Theorem 2.1 with the convexity assumption replaced by the assumption $\delta(\mathcal{H}) := \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le K\epsilon$. In this section we will develop further the idea of replacing the convexity assumption by an estimate for $\delta(\mathcal{H})$. The motivation for this is that applications of results of the type of Theorem 3.1 in construction of universal estimators require bounds in a more general situation than $\delta(\mathcal{H}) \le K\epsilon$. The following theorem provides bounds for $\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \epsilon\}$ in the case of arbitrary ϵ and $\delta(\mathcal{H})$.

Theorem 3.2. *Suppose $\mathcal{H}$ is a compact subset of $\mathcal{C}(X)$ and $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le \delta$. Assume that ρ, $\mathcal{H}$ satisfy (1.5). Then, for all $\epsilon > 0$*

$$\rho^m\{\mathbf{z} : \mathcal{E}(f_{\mathbf{z},\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \epsilon\} \le N(\mathcal{H}, \frac{\epsilon}{16M}, \mathcal{C}(X))\exp\big(-\frac{m\epsilon^2}{2^9M^2(\epsilon+\delta)}\big).$$

Lemma 3.1 [DKPT1]. *For any f we have*

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

Proof. We have

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

Next,

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

Combining the above two relations we get

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

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

Lemma 3.2. *Assume that ρ and $\mathcal{H}$ satisfy (1.5). Then we have*

$$\sigma^2 := \sigma^2(\ell(f)) \le 8M^2(\mathcal{E}_{\mathcal{H}}(f) + 2\delta(\mathcal{H})).$$

Proof. We have

$$\sigma^2(\ell(f)) \le E(\ell(f)^2) = E((f(x) - f_{\mathcal{H}}(x))^2(f(x) + f_{\mathcal{H}}(x) - 2y)^2)$$

$$\le 4M^2 E((f(x) - f_{\mathcal{H}}(x))^2) = 4M^2\|f - f_{\mathcal{H}}\|^2_{L_2(\rho_X)} \le 8M^2(\mathcal{E}_{\mathcal{H}}(f) + 2\delta(\mathcal{H})).$$

At the last step we have used Lemma 3.1. □

Lemma 3.3. *Assume that ρ and $\mathcal{H}$ satisfy (1.5). Let $f \in \mathcal{H}$. For all $\epsilon > 0$, $\alpha \in (0, 1]$ one has*

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \ge \alpha(\mathcal{E}_{\mathcal{H}}(f) + \epsilon)\} \le \exp\Big(-\frac{\alpha^2 m\epsilon^2}{32M^2(\epsilon + \delta(\mathcal{H}))}\Big).$$

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \le -\alpha(\mathcal{E}_{\mathcal{H}}(f) + \epsilon)\} \le \exp\Big(-\frac{\alpha^2 m\epsilon^2}{32M^2(\epsilon + \delta(\mathcal{H}))}\Big).$$

Proof. Denote $a := \mathcal{E}_{\mathcal{H}}(f)$. The proofs of both inequalities are the same. We will carry on only the proof of the first one. Using one-sided Bernstein's inequality (1.8) for $\ell(f)$ we obtain

$$\rho^m\{\mathbf{z} : \mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) \ge \alpha(a + \epsilon)\} \le \exp\Big(-\frac{m\alpha^2(a + \epsilon)^2}{2(\sigma^2 + M^2\alpha(a + \epsilon)/3)}\Big).$$

It remains to check that

$$\frac{(a + \epsilon)^2}{2(\sigma^2 + M^2\alpha(a + \epsilon)/3)} \ge \frac{\epsilon^2}{32M^2(\epsilon + \delta(\mathcal{H}))}. \tag{3.1}$$

Using Lemma 3.2 we get on the one hand

$$\epsilon^2(\sigma^2 + M^2\alpha(a + \epsilon)/3) \le M^2\epsilon^2(9a + 16\delta(\mathcal{H}) + \epsilon/3). \tag{3.2}$$

On the other hand

$$16M^2(\epsilon + \delta(\mathcal{H}))(a + \epsilon)^2 \ge M^2\epsilon^2(32a + 16\delta(\mathcal{H}) + 16\epsilon). \tag{3.3}$$

Comparing (3.2) and (3.3) we obtain (3.1). □

A combination of Lemma 3.3 and Lemma 2.4 gives the following theorem.

Theorem 3.3. *Assume that ρ and $\mathcal{H}$ satisfy (1.5) and are such that $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le \delta$. Then for all $\epsilon > 0$ and $\alpha \in (0,1)$*

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f)+\epsilon} \ge 2\alpha\} \le N(\mathcal{H}, \frac{\alpha\epsilon}{4M}, \mathcal{C}(X)) \exp\big(-\frac{\alpha^2 m\epsilon^2}{32M^2(\epsilon+\delta)}\big).$$

Proof. Let $f_1, \dots, f_N$ be the $\frac{\alpha\epsilon}{4M}$-net of $\mathcal{H}$ in $\mathcal{C}(X)$, $N := N(\mathcal{H}, \frac{\alpha\epsilon}{4M}, \mathcal{C}(X))$. Let Λ be the set of $\mathbf{z}$ such that for all $j = 1, \dots, N$ we have

$$\frac{\mathcal{E}_{\mathcal{H}}(f_j) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f_j)}{\mathcal{E}_{\mathcal{H}}(f_j)+\epsilon} < \alpha.$$

Then by Lemma 3.3

$$\rho^m(\Lambda) \ge 1 - N \exp\big(-\frac{\alpha^2 m\epsilon^2}{32M^2(\epsilon+\delta)}\big). \tag{3.4}$$

We take any $\mathbf{z} \in \Lambda$ and any $g \in \mathcal{H}$. Let f_j be such that $\|g - f_j\|_{\mathcal{C}(X)} \le \frac{\alpha\epsilon}{4M}$. By Lemma 2.4 we obtain that

$$\frac{\mathcal{E}_{\mathcal{H}}(g) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathcal{H}}(g)+\epsilon} < 2\alpha.$$

It remains to use (3.4). □

Theorem 3.4. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$ such that $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \le \delta$. Assume that ρ, $\mathcal{H}$ satisfy (1.5). Then for all $\epsilon > 0$ with probability at least*

$$p(\mathcal{H}, \epsilon, \delta) := 1 - N(\mathcal{H}, \frac{\epsilon}{16M}, \mathcal{C}(X)) \exp\big(-\frac{m\epsilon^2}{2^9 M^2(\epsilon+\delta)}\big)$$

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

$$\mathcal{E}(f) \le 2\mathcal{E}_{\mathbf{z}}(f) + \epsilon - \mathcal{E}(f_{\mathcal{H}}) + 2(\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})). \tag{3.5}$$

Proof. Using Theorem 3.3 with $\alpha = 1/4$ we get with probability at least $p(\mathcal{H}, \epsilon, \delta)$

$$\mathcal{E}_{\mathcal{H}}(f) \le 2\mathcal{E}_{\mathcal{H},\mathbf{z}}(f) + \epsilon. \tag{3.6}$$

Substituting

$$\mathcal{E}_{\mathcal{H}}(f) := \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}); \qquad \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})$$

we obtain (3.5). □

Corollary 3.1. *Under the assumptions of Theorem 3.4 we have*

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}}) \le \epsilon/2$$

with probability at least $p(\mathcal{H}, \epsilon, \delta)$.

Proof of Theorem 3.2. The statement of the theorem follows immediately from (3.6) with $f = f_{\mathbf{z},\mathcal{H}}$ because $\mathcal{E}_{\mathcal{H},\mathbf{z}}(f_{\mathbf{z},\mathcal{H}}) \le 0$ from the definition of $f_{\mathbf{z},\mathcal{H}}$.

Theorem 3.5. *Let* W *be a compact in* $L_1(\rho_X)$ *set and let* ρ, W *satisfy (1.5). Assume* $W \in \mathcal{S}_1^r$ *that is*

$$\epsilon_n(W, L_1(\rho_X)) \le Dn^{-r}, \quad n = 1, 2, \dots, \quad W \subset DU(L_1(\rho_X)). \tag{3.7}$$

Then there exists an estimator $f_{\mathbf{z}}$ *such that for* $\eta \ge \eta_m := (6M+4)\epsilon_0$,

$$\epsilon_0 := C(M, D, r)\big(m^{-1}\max(m^{-\frac{r}{1+r}}, \delta(W))\big)^{\frac{r}{1+2r}}, \quad m \ge C_1(M, D, r),$$

we have

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

Proof. Let $\mathcal{N} := \mathcal{N}_{\epsilon_0}(W, L_1(\rho_X))$ be a minimal ϵ_0-net of W in the $L_1(\rho_X)$ norm. The constant $C(M, D, r)$ will be chosen later. Then (3.7) implies that

$$|\mathcal{N}| \le 2^{(D/\epsilon_0)^{1/r}+1}. \tag{3.8}$$

As an estimator $f_{\mathbf{z}}$ we take

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

We take $\epsilon \ge \epsilon_0$ and apply the first inequality of Lemma 3.3 with $\alpha = 1/2$ to each $f \in \mathcal{N}$. In such a way we obtain a set Λ_1 with

$$\rho^m(\Lambda_1) \ge 1 - |\mathcal{N}| \exp\big(-\frac{m\epsilon^2}{128M^2(\epsilon + \delta(W))}\big)$$

with the property: for all $f \in \mathcal{N}$ and all $\mathbf{z} \in \Lambda_1$ one has

$$\mathcal{E}_W(f) \le 2\mathcal{E}_{W,\mathbf{z}}(f) + \epsilon. \tag{3.9}$$

Therefore, for $\mathbf{z} \in \Lambda_1$

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathbf{z}}) + \epsilon \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) + \epsilon. \tag{3.10}$$

Let Λ_2 be the set of all $\mathbf{z}$ such that

$$\mathcal{E}_W(f_{\mathcal{N}}) - \mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) \le -\frac{1}{2}(\mathcal{E}_W(f_{\mathcal{N}}) + \epsilon). \tag{3.11}$$

By the second inequality of Lemma 3.3 with $\alpha = 1/2$

$$\rho^m(\Lambda_2) \le \exp\big(-\frac{m\epsilon^2}{128M^2(\epsilon + \delta(W))}\big).$$

Consider $\Lambda := \Lambda_1 \setminus \Lambda_2$. Then

$$\rho^m(\Lambda) \ge 1 - (|\mathcal{N}| + 1)\exp\big(-\frac{m\epsilon^2}{128M^2(\epsilon + \delta(W))}\big).$$

Using the inequality opposite to (3.11) we continue (3.10) for $\mathbf{z} \in \Lambda$

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 2\mathcal{E}_{W,\mathbf{z}}(f_{\mathcal{N}}) + \epsilon \le 3\mathcal{E}_W(f_{\mathcal{N}}) + 2\epsilon.$$

By (2.17) we obtain from here

$$\mathcal{E}_W(f_{\mathbf{z}}) \le 6M\epsilon_0 + 2\epsilon.$$

We choose $\epsilon_0 := C(M, D, r)\big(m^{-1}\max(m^{-\frac{r}{1+r}}, \delta(W))\big)^{\frac{r}{1+2r}}$ from the inequality

$$3(D/\epsilon_0)^{\frac{1}{r}} - \frac{m\epsilon_0^2}{128M^2(\epsilon_0 + \delta(W))} \le 0.$$

Then for $m \ge C_1(M, D, r)$ we have $\epsilon_0 \le D$. We let $\eta = 6M\epsilon_0 + 2\epsilon$. Then our assumption $\eta \ge (6M + 4)\epsilon_0$ implies $\epsilon \ge 2\epsilon_0$. Using the inequality

$$\frac{\epsilon_0 + \delta}{\epsilon + \delta} \ge \frac{\epsilon_0}{\epsilon}, \quad \epsilon \ge \epsilon_0,$$

we obtain

$$\frac{\epsilon^2}{\epsilon + \delta(W)} - \frac{\epsilon_0^2}{\epsilon_0 + \delta(W)} \ge \frac{1}{2}\frac{\epsilon^2}{\epsilon + \delta(W)}.$$

Therefore

$$\rho^m(Z^m \setminus \Lambda)$$

$$\le (|\mathcal{N}| + 1)\exp\big(-\frac{m\epsilon_0^2}{128M^2(\epsilon_0 + \delta(W))}\big)\exp\big(-\frac{m\epsilon^2}{128M^2(\epsilon + \delta(W))} + \frac{m\epsilon_0^2}{128M^2(\epsilon_0 + \delta(W))}\big)$$

$$\le \exp\big(-\frac{m\epsilon^2}{256M^2(\epsilon + \delta(W))}\big) \le \exp\big(-\frac{c(M)m\eta^2}{\eta + \delta(W)}\big).$$

This completes the proof of Theorem 3.5. □

Let a, b, be two positive numbers. Consider a collection $\mathcal{K}(a, b)$ of compacts K_n in $\mathcal{C}(X)$ satisfying

$$N(K_n, \epsilon, \mathcal{C}(X)) \le (a(1 + 1/\epsilon))^n n^{bn}, \quad n = 1, 2, \dots. \tag{3.12}$$

Denote $\delta_j := Aj^{-2r}$, $\epsilon_j := \frac{Aj\ln m}{m}$. Let j_r be the minimal $j \in (1, m]$ such that $\epsilon_j \ge \delta_j$. Then

$$j_r \asymp \big(\frac{m}{\ln m}\big)^{\frac{1}{1+2r}}.$$

Denote $\epsilon(r) := \epsilon_{j_r}$.

Lemma 3.4. *Assume $\{K_n\}_{n=1}^m$ satisfy (3.12) and $r \le 1/2$. Then for $A \ge C(a,b,M)$*

$$\sum_{j=j_r}^{m}(1-p(K_j,\epsilon_j,\delta_j)) + \sum_{j<j_r}(1-p(K_j,\epsilon(r),\delta_j)) \le m^{-c(M)A}.$$

Proof. For estimating $1-p(K_j,\epsilon_j,\delta_j)$, $j \in [j_r,m]$ we write

$$\ln\big(N(K_j,\epsilon_j/(16M),\mathcal{C}(X))\exp\big(-\frac{m\epsilon_j^2}{2^9M^2(\epsilon_j+\delta_j)}\big)\big) \tag{3.13}$$

$$\le j\ln(a(1+16M/\epsilon_j)) + bj\ln j - \frac{m\epsilon_j}{2^{10}M^2} \le j\ln(a(1+16M)) + j(1+b)\ln m - Ac_2(M)j\ln m$$

$$\le -Ac_3(M)\ln m$$

for $A \ge C_1(a,b,M)$.

We proceed to the case $j < j_r$. We now have

$$\ln\big(N(K_j,\epsilon(r)/(16M),\mathcal{C}(X))\exp\big(-\frac{m\epsilon(r)^2}{2^9M^2(\epsilon(r)+\delta_j)}\big)\big) \tag{3.14}$$

$$\le j\ln(a(1+16M/\epsilon(r))) + bj\ln j - \frac{m\epsilon(r)^2}{2^{10}M^2\delta_j} \le c_2(a,b,M)j\ln m - Ac_5(M)m\big(\frac{\ln m}{m}\big)^{\frac{4r}{1+2r}}j^{2r}$$

$$\le j\big(c_2(a,b,M)\ln m - Ac_5(M)m\big(\frac{\ln m}{m}\big)^{\frac{4r}{1+2r}}j_r^{2r-1}\big) \le -Ac_6(M)\ln m$$

for $A \ge C_2(a,b,M)$.

Combining (3.13) and (3.14) we complete the proof of Lemma 3.4. □

Lemma 3.4 allows us to use the inequality (3.5) simultaneously for all $j = 1,\dots,m$ with $\mathcal{H} = K_j$, $\epsilon = \max(\epsilon_j,\epsilon(r))$, $\delta = \delta_j$.

Theorem 3.6. *Let compacts K_j, $j = 1,\dots,m$, satisfy (3.12) and all pairs ρ, K_j, $j = 1,\dots,m$, satisfy (1.5). Assume that*

$$\mathcal{E}(f_{K_j}) - \mathcal{E}(f_\rho) \le Aj^{-2r}, \quad r \in (0,1/2].$$

Then there exists $A_0(a,b,M)$ such that for $A \ge A_0(a,b,M)$ we have all the inequalities

$$\mathcal{E}_{K_j}(f) \le 2\mathcal{E}_{K_j,\mathbf{z}}(f) + \frac{A\max(j,j_r)\ln m}{m}, \quad f \in K_j, \quad j \in [1,m], \tag{3.15}$$

where j_r is the minimal $j \in (1,m]$ such that $j^{1+2r} \ge m/\ln m$, with probability $\ge 1 - m^{-c(M)A}$.

Proof. By Theorem 3.4 (see (3.6)) with $\mathcal{H} = K_j$, $\epsilon = \frac{A\max(j,j_r)\ln m}{m}$, $\delta = Aj^{-2r}$ we get (3.15) for $j \in [1,m]$ with the probability

$$p \ge 1 - \Big(\sum_{j=j_r}^{m}(1-p(K_j,\epsilon_j,\delta_j)) + \sum_{j<j_r}(1-p(K_j,\epsilon(r),\delta_j))\Big) \ge 1 - m^{-c(M)A}$$

with the ϵ_j, δ_j, $\epsilon(r)$ defined above. It remains to use Lemma 3.4. □

Corollary 3.2. *Under the assumptions of Theorem 3.6 we have the inequalities*

$$\mathcal{E}_{\mathbf{z}}(f_{K_j}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},K_j}) \le \frac{A \max(j, j_r) \ln m}{2m}, \quad j \in [1, m], \tag{3.16}$$

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

Proof. The inequalities (3.16) follows from (3.15) with $f = f_{\mathbf{z},K_j}$ if we note that $\mathcal{E}_{K_j}(f) \ge 0$, $f \in K_j$. $\square$

4. Penalized least squares estimators

The technique from Section 2 can also be used in the following situation. Define

$$\mathcal{E}_\rho(f) := \mathcal{E}(f) - \mathcal{E}(f_\rho); \qquad \mathcal{E}_{\rho,\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho);$$

$$\ell_\rho(f) := (f(x) - y)^2 - (f_\rho(x) - y)^2.$$

Then we have the following analogues of the lemmas from Section 2.

Lemma 4.1. *Let f, f_ρ be such that $|f(x) - y| \le M$, $|f_\rho(x) - y| \le M$ a.s. Then we have*

$$\sigma^2(\ell_\rho(f)) \le 4M^2 \mathcal{E}_\rho(f).$$

Lemma 4.2. *Let f, f_ρ be such that $|f(x) - y| \le M$, $|f_\rho(x) - y| \le M$ a.s. Then for all $\epsilon > 0$, $\alpha \in (0, 1]$ one has*

$$\rho^m\{\mathbf{z} : \mathcal{E}_\rho(f) - \mathcal{E}_{\rho,\mathbf{z}}(f) \ge \alpha(\mathcal{E}_\rho(f) + \epsilon)\} \le \exp\Big(-\frac{\alpha^2 m\epsilon}{5M^2}\Big),$$

$$\rho^m\{\mathbf{z} : \mathcal{E}_\rho(f) - \mathcal{E}_{\rho,\mathbf{z}}(f) \le -\alpha(\mathcal{E}_\rho(f) + \epsilon)\} \le \exp\Big(-\frac{\alpha^2 m\epsilon}{5M^2}\Big).$$

Lemma 4.3. *Assume ρ, $\mathcal{H}$ satisfy (1.5) and $f, g \in \mathcal{H}$ are such that $\|f - g\|_{\mathcal{C}(X)} \le \alpha\epsilon/(4M)$. Let $\alpha \in (0, 1)$, $\epsilon > 0$, and let f be such that*

$$\frac{\mathcal{E}_\rho(f) - \mathcal{E}_{\rho,\mathbf{z}}(f)}{\mathcal{E}_\rho(f) + \epsilon} < \alpha.$$

Then we have

$$\frac{\mathcal{E}_\rho(g) - \mathcal{E}_{\rho,\mathbf{z}}(g)}{\mathcal{E}_\rho(g) + \epsilon} < 2\alpha.$$

These lemmas imply the following analogue of Theorem 2.3.

Theorem 4.1. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that ρ and $\mathcal{H}$ satisfy (1.5). Then for all $\epsilon > 0$ with probability at least*

$$p(\mathcal{H}, \rho, \epsilon) := 1 - N(\mathcal{H}, \frac{\epsilon}{16M}, \mathcal{C}(X)) \exp\big(-\frac{m\epsilon}{80M^2}\big)$$

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

$$\mathcal{E}_\rho(f) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f) + \epsilon. \tag{4.1}$$

We first demonstrate how Lemma 4.2 can be used in proving optimal upper estimates.

Theorem 4.2. *Let $f_\rho \in \Theta$ and let ρ, Θ satisfy (1.5). Assume*

$$\epsilon_n(\Theta, L_2(\rho_X)) \le Dn^{-r}, \quad n = 1, 2, \dots, \quad \Theta \subset DU(L_2(\rho_X)). \tag{4.2}$$

Then there exists an estimator $f_\mathbf{z}$ such that for $\eta \ge \eta_m := 7\epsilon_0$, $\epsilon_0 := C(M, D, r)m^{-\frac{2r}{1+2r}}$, $m \ge 60(M/D)^2$, we have

$$\rho^m\{\mathbf{z} : \|f_\mathbf{z} - f_\rho\|^2_{L_2(\rho_X)} \ge \eta\} \le \exp\big(-\frac{m\eta}{200M^2}\big).$$

Proof. Let $\mathcal{N} := \mathcal{N}_{\epsilon_0^{1/2}}(\Theta, L_2(\rho_X))$ be a minimal $\epsilon_0^{1/2}$-net of Θ in the $L_2(\rho_X)$ norm. The constant $C(M, D, r)$ will be chosen later. Then (4.2) implies that

$$|\mathcal{N}| \le 2^{(D^2/\epsilon_0)^{1/(2r)}+1}. \tag{4.3}$$

As an estimator $f_\mathbf{z}$ we take

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

We take $\epsilon \ge \epsilon_0$ and apply the first inequality of Lemma 4.2 with $\alpha = 1/2$ to each $f \in \mathcal{N}$. In such a way we obtain a set Λ_1 with

$$\rho^m(\Lambda_1) \ge 1 - |\mathcal{N}| \exp\big(-\frac{m\epsilon}{20M^2}\big)$$

with the property: for all $f \in \mathcal{N}$ and all $\mathbf{z} \in \Lambda_1$ one has

$$\mathcal{E}_\rho(f) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f) + \epsilon. \tag{4.4}$$

Therefore, for $\mathbf{z} \in \Lambda_1$

$$\mathcal{E}_\rho(f_\mathbf{z}) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f_\mathbf{z}) + \epsilon \le 2\mathcal{E}_{\rho,\mathbf{z}}(f_\mathcal{N}) + \epsilon. \tag{4.5}$$

Let Λ_2 be the set of all $\mathbf{z}$ such that

$$\mathcal{E}_\rho(f_\mathcal{N}) - \mathcal{E}_{\rho,\mathbf{z}}(f_\mathcal{N}) \le -\frac{1}{2}(\mathcal{E}_\rho(f_\mathcal{N}) + \epsilon). \tag{4.6}$$

By the second inequality of Lemma 4.2 with $\alpha = 1/2$

$$\rho^m(\Lambda_2) \le \exp\big(-\frac{m\epsilon}{20M^2}\big).$$

Consider $\Lambda := \Lambda_1 \setminus \Lambda_2$. Then

$$\rho^m(\Lambda) \ge 1 - (|\mathcal{N}| + 1)\exp\big(-\frac{m\epsilon}{20M^2}\big).$$

Using the inequality opposite to (4.6) we continue (4.5) for $\mathbf{z} \in \Lambda$

$$\mathcal{E}_\rho(f_\mathbf{z}) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f_\mathcal{N}) + \epsilon \le 3\mathcal{E}_\rho(f_\mathcal{N}) + 2\epsilon \le 3\epsilon_0 + 2\epsilon.$$

We choose $\epsilon_0 \le D^2$ from the equation

$$3(D^2/\epsilon_0)^{\frac{1}{2r}} = \frac{m\epsilon_0}{20M^2}.$$

We get

$$\epsilon_0 = (60M^2)^{\frac{2r}{1+2r}} D^{\frac{2}{1+2r}} m^{-\frac{2r}{1+2r}}.$$

For $m \ge 60(M/D)^2$ we have $\epsilon_0 \le D^2$. We let $\eta = 3\epsilon_0 + 2\epsilon$. Then our assumption $\eta \ge 7\epsilon_0$ implies $\epsilon \ge 2\epsilon_0$ and

$$\rho^m(Z^m \setminus \Lambda) \le (|\mathcal{N}| + 1)\exp\big(-\frac{m\epsilon_0}{20M^2}\big)\exp\big(-\frac{m(\epsilon - \epsilon_0)}{20M^2}\big) \le \exp\big(-\frac{m\epsilon}{40M^2}\big) \le \exp\big(-\frac{m\eta}{200M^2}\big).$$

This completes the proof of Theorem 4.2. □

Let us compare Theorem 4.2 with Theorem 3.5 in the case $\delta(W) = 0$, what corresponds to the assumption $f_\rho \in W$. We note that assumptions $\Theta \in \mathcal{S}_1^r$ and $\Theta \subset MU(L_\infty(\rho_X))$ imply that

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

Therefore, for $\Theta \in \mathcal{S}_1^r$, $\Theta \subset MU(L_\infty(\rho_X))$ Theorem 4.2 implies

$$\eta_m \ll m^{-\frac{2(r/2)}{1+2(r/2)}} = m^{-\frac{r}{1+r}}$$

which corresponds to the estimate from Theorem 3.5. Thus, in the case $f_\rho \in \Theta$ Theorem 4.2 implies Theorem 3.5.

We proceed to the universal estimators. Let as above $\mathcal{K} := \mathcal{K}(a, b)$ be a collection of compacts K_n in $\mathcal{C}(X)$ satisfying (3.12). We take a parameter $A \ge 1$ and consider the following estimator

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

with

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

Theorem 4.3. *For $\mathcal{K} := \{K_n\}_{n=1}^{\infty}$ satisfying (3.12) there exists $A_0 := A_0(a,b,M)$ such that for any $A \ge A_0$ and any ρ such that ρ, K_n, $n = 1,2,\dots$ satisfy (1.5) we have*

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

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

Proof. We set $\epsilon_j := \frac{2Aj\ln m}{m}$ for all $j \in [1,m]$. Applying Theorem 4.1 and Lemma 4.2 we find a set Λ with

$$\rho^m(\Lambda) \ge 1 - m^{-c(M)A}$$

such that for all $\mathbf{z} \in \Lambda$, $j \in [1,m]$ we have

$$\mathcal{E}_\rho(f) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f) + \epsilon_j, \quad \forall f \in K_j,$$

$$\mathcal{E}_{\rho,\mathbf{z}}(f_{K_j}) \le \frac{3}{2}\mathcal{E}_\rho(f_{K_j}) + \epsilon_j/2.$$

We get from here for $\mathbf{z} \in \Lambda$

$$\mathcal{E}_\rho(f_{\mathbf{z}}^A) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z},K_{n(\mathbf{z})}}) + \epsilon_{n(\mathbf{z})} = 2\big(\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z},K_{n(\mathbf{z})}}) + \frac{An(\mathbf{z})\ln m}{m}\big)$$

$$= 2\min_{1\le j\le m}\big(\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z},K_j}) + \frac{Aj\ln m}{m}\big) \le 2\min_{1\le j\le m}\big(\mathcal{E}_{\rho,\mathbf{z}}(f_{K_j}) + \frac{Aj\ln m}{m}\big)$$

$$\le 2\min_{1\le j\le m}\big(\frac{3}{2}\mathcal{E}_\rho(f_{K_j}) + \frac{2Aj\ln m}{m}\big). \qquad \square$$

We note that results in a style of Theorem 4.3 with bounds of the expectation $E_{\rho^m}(\|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)}^2)$ are known (see [GKKW, Ch.12]).

Theorem 4.4. *Let compacts $\{K_n\}$ satisfy (3.12). There exists $A_0 := A_0(a,b,M) \ge 1$ such that for any $A \ge A_0$ and any ρ satisfying*

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

and such that ρ, K_n, $n = 1,2,\dots$, satisfy (1.5) we have for $\eta \ge A^{1/2}\big(\frac{\ln m}{m}\big)^{\frac{r}{1+2r}}$

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

Proof. Let $\eta \ge A^{1/2}(\frac{\ln m}{m})^{\frac{r}{1+2r}}$. We define n as the smallest integer such that $2n \ge m\eta^2/\ln m$. Denote $\epsilon_j := \frac{2Aj\ln m}{m}$, $j \in (n,m]$; $\epsilon_j := A\eta^2$, $j \in [1,n]$. We apply Theorem 4.1 to K_j with $\epsilon = \epsilon_j$. Denote Λ_j the set of all $\mathbf{z}$ such that for any $f \in K_j$

$$\mathcal{E}_\rho(f) \le 2\mathcal{E}_{\rho,\mathbf{z}}(f) + \epsilon_j. \tag{4.7}$$

By Theorem 4.1

$$\rho^m(\Lambda_j) \ge p(K_j, \rho, \epsilon_j).$$

For estimating $p(K_j, \rho, \epsilon_j)$ we write ($j \in [n, m]$)

$$\ln\big(N(K_j, \epsilon_j/(16M), \mathcal{C}(X)) \exp\big(-\frac{m\epsilon_j}{80M^2}\big)\big)$$

$$\le j \ln(a(1 + 16M/\epsilon_j)) + bj \ln j - \frac{m\epsilon_j}{80M^2} \le j \ln(a(1 + 8M)) + j(1 + b) \ln m - Ac_2(M) j \ln m$$

$$\le -Ac_3(M) j \ln m \le -Ac_3(M) n \ln m \le -Ac_4(M) m\eta^2$$

for $A \ge C_1(a, b, M)$. Similar estimate for $j \in [1, n]$ follows from the above estimate with $j = n$.

Thus (4.7) holds for all $1 \le j \le m$ on the set $\Lambda' := \cap_{j=1}^m \Lambda_j$ with

$$\rho^m(\Lambda') \ge 1 - e^{-c_5(M) m\eta^2}. \tag{4.8}$$

For $j \in [1, m]$ we have by the assumption of Theorem 4.4 that

$$\mathcal{E}_\rho(f_{K_j}) = \|f_{K_j} - f_\rho\|^2_{L_2(\rho_X)} \le Aj^{-2r}. \tag{4.9}$$

We apply the second inequality of Lemma 4.2 to each f_{K_j} with $\alpha = 1/2$ and ϵ_j chosen above, $j = 1, \dots, m$. Then we obtain a set Λ'' of $\mathbf{z}$ such that

$$\mathcal{E}_{\rho,\mathbf{z}}(f_{K_j}) \le \frac{3}{2}\mathcal{E}_\rho(f_{K_j}) + \epsilon_j/2, \quad j = 1, \dots, m, \tag{4.10}$$

and

$$\rho^m(\Lambda'') \ge 1 - \sum_{j=1}^m \exp\big(-\frac{m\epsilon_j}{20M^2}\big) \ge 1 - e^{-c_6(M) m\eta^2}. \tag{4.11}$$

For the set $\Lambda := \Lambda' \cap \Lambda''$ we have the inequalities (4.7) and (4.10) for all $j \in [1, m]$. Let $\mathbf{z} \in \Lambda$. We apply (4.7) to $f^A_{\mathbf{z}} = f_{\mathbf{z}, K_{n(\mathbf{z})}}$. We consider separately two cases: I. $n(\mathbf{z}) > n$; II. $n(\mathbf{z}) \le n$. In the first case we obtain

$$\mathcal{E}_\rho(f^A_{\mathbf{z}}) \le 2\big(\mathcal{E}_{\rho,\mathbf{z}}(f^A_{\mathbf{z}}) + \frac{An(\mathbf{z}) \ln m}{m}\big). \tag{4.12}$$

Using the definition of $f^A_{\mathbf{z}}$ and the inequality (4.10) we get

$$\mathcal{E}_{\rho,\mathbf{z}}(f^A_{\mathbf{z}}) + \frac{An(\mathbf{z}) \ln m}{m} = \min_{1 \le j \le m} \big(\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z},K_j}) + \frac{Aj \ln m}{m}\big) \tag{4.13}$$

$$\le \min_{1\le j\le m}\big(\mathcal{E}_{\rho,\mathbf{z}}(f_{K_j})+\frac{Aj\ln m}{m}\big)$$

$$\le \min\big(\min_{n<j\le m}\big(\frac{3}{2}\mathcal{E}(f_{K_j})+\frac{2Aj\ln m}{m}\big),\min_{1\le j\le n}\big(\frac{3}{2}\mathcal{E}(f_{K_j})+\frac{Aj\ln m}{m}\big)+A\eta^2/2\big)$$

$$\le \min_{1\le j\le m}\big(\frac{3}{2}\mathcal{E}(f_{K_j})+\frac{2Aj\ln m}{m}\big)+A\eta^2/2\le \min_{1\le j\le m}\big(\frac{3}{2}Aj^{-2r}+\frac{2Aj\ln m}{m}\big)+A\eta^2/2.$$

Substituting $j=[(m/\ln m)^{\frac{1}{1+2r}}]+1$ and using the inequalities

$$(m/\ln m)^{\frac{1}{1+2r}}\le j\le 2(m/\ln m)^{\frac{1}{1+2r}}$$

we obtain from (4.12) and (4.13)

$$\mathcal{E}_\rho(f_{\mathbf{z}}^A)\le 11A\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}}+A\eta^2\le 12A\eta^2.$$

This gives the required bound

$$\|f_{\mathbf{z}}^A-f_\rho\|_{L_2(\rho_X)}\le 4A^{1/2}\eta.$$

In the second case we obtain

$$\mathcal{E}_\rho(f_{\mathbf{z}}^A)\le 2\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z}}^A)+A\eta^2.$$

Next, we have

$$\mathcal{E}_{\rho,\mathbf{z}}(f_{\mathbf{z}}^A)\le \min_{1\le j\le m}\big(\mathcal{E}_{\rho,\mathbf{z}}(f_{K_j})+\frac{Aj\ln m}{m}\big).$$

Using Lemma 4.2 we continue

$$\le \min_{1\le j\le m}\big(\frac{3}{2}\mathcal{E}_\rho(f_{K_j})+A\eta^2/2+\frac{2Aj\ln m}{m}\big)$$

$$\le \min_{1\le j\le m}\big(\frac{3}{2}Aj^{-2r}+\frac{2Aj\ln m}{m}\big)+A\eta^2/2\le 6A\eta^2.$$

Therefore,

$$\mathcal{E}_\rho(f_{\mathbf{z}}^A)\le 13A\eta^2.$$

The proof of Theorem 4.2 is complete. □

As it is mentioned in the Introduction in [DKPT2] we proposed to study the **AC**-function. In the discussion that follows it will be more convenient for us to express the results in terms of the following variant of the accuracy confidence function

$$\mathbf{ac}_m(\mathcal{M},\mathbb{E},\eta):=\mathbf{AC}_m(\mathcal{M},\mathbb{E},\eta^{1/2}).$$

We may study the **ac**-function in two steps.

Step 1. For given $\mathcal{M}$, m, $E(m)$ find for $\delta \in (0,1)$ the smallest $t_m(\mathcal{M},\delta) := t_m(\mathcal{M},E(m),\delta)$ such that

$$\mathbf{ac}_m(\mathcal{M},\mathbb{E},t_m(\mathcal{M},\delta)) \le \delta.$$

It is clear that for $\eta > t_m(\mathcal{M},\delta)$ we have $\mathbf{ac}_m(\mathcal{M},\mathbb{E},\eta) \le \delta$ and for $\eta < t_m(\mathcal{M},\delta)$ we have $\mathbf{ac}_m(\mathcal{M},\mathbb{E},\eta) > \delta$.

The following modification of the above $t_m(\mathcal{M},\delta)$ is also of interest. We now look for the smallest $t_m(\mathcal{M},\delta,c)$ such that

$$\mathbf{ac}_m(\mathcal{M},\mathbb{E},t_m(\mathcal{M},\delta,c)) \le \delta m^{-c}, \quad c > 0.$$

It is clear that $t_m(\mathcal{M},\delta) \le t_m(\mathcal{M},\delta,c)$. We call the $t_m(\mathcal{M},\delta)$ and $t_m(\mathcal{M},\delta,c)$ the *approximation threshold for the proper function learning*.

Step 2. Find the right order of $\mathbf{ac}_m(\mathcal{M},\mathbb{E},\eta)$ for $\eta \ge t_m(\mathcal{M},\delta)$ as a function on m and η.

It has been proved in [DKPT2], [T2] (see Theorem 1.2 of Introduction) that for a compact $\Theta \subset L_2(\mu)$ such that $\Theta \subset \frac{1}{4}U(L_\infty(\mu))$ and

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}, \tag{4.14}$$

one has: there exists $\delta_0 > 0$ such that for any $\delta \in (0,\delta_0]$

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

Theorem 4.2 implies that under the above assumptions

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

Therefore, for any Θ satisfying the above assumptions we have

$$t_m(\mathcal{M}(\Theta,\mu),\delta) \asymp m^{-\frac{2r}{1+2r}}, \quad \delta \in (0,\delta_0]. \tag{4.15}$$

We now proceed to the concept of universal (universally optimal) estimator. Let a collection $\mathbb{M} := \{\mathcal{M}\}$ of classes $\mathcal{M}$ of measures and a sequence $\mathbb{E}$ of allowed classes of estimators be given.

Definition 4.1. *An estimator $f_{\mathbf{z}} \in E(m)$ is a universal (universally optimal) in a weak sense for the pair $(\mathbb{M},\mathbb{E})$ if for any $\rho \in \mathcal{M} \in \mathbb{M}$ we have*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|^2_{L_2(\rho_X)} \ge C_1(\mathbb{M},\mathbb{E})(\ln m)^w t_m(\mathcal{M},\delta,c)\} \le C_2 m^{-c_1},$$

where C_1, c_1, and C_2 do not depend on ρ and m.

In the case $w = 0$ in the above definition we replace *in a weak sense* by *in a strong sense.*

We now discuss an application of Theorem 4.4 for construction of universal estimators. Let $\mathcal{L} := \{L_n\}_{n=1}^{\infty}$ be a sequence of n-dimensional subspaces of $\mathcal{C}(X)$. Consider the $\mathcal{C}(X)$-balls in L_n of radius D:

$$K_n := DU(\mathcal{C}(X)) \cap L_n, \quad n = 1, 2, \dots .$$

Then the sequence $\mathcal{K} := \{K_n\}$ satisfy (3.12) with $a = 2D$. Consider the classes

$$\Theta^r(\mathcal{K}, \mu) := \{f : d(f, K_n)_{L_2(\mu)} \le C_3 n^{-r}, \quad n = 1, 2, \dots\}, \tag{4.16}$$

with C_3 a fixed positive number. We also consider a set V of Borel probability measures ν defined on X such that

$$\epsilon_n(\Theta^r(\mathcal{K}, \nu), L_2(\nu)) \ge C_4 n^{-r}, \quad \nu \in V, \quad r \in [\alpha, \beta]. \tag{4.17}$$

We consider a class $\mathcal{M}(r, \nu)$ of measures ρ such that $|y| \le M$ a.e. with respect to ρ and $\rho_X = \nu$, $f_\rho \in \Theta^r(\mathcal{K}, \nu)$. Finally, we define a collection

$$\mathbb{M} := \{\rho : \rho \in \mathcal{M}(r, \nu), \quad \nu \in V, \quad r \in [\alpha, \beta]\}.$$

Then for any $\nu \in V$, $r \in [\alpha, \beta]$ our assumptions (4.16) and (4.17) imply (by Carl's inequality [C]) that

$$\epsilon_n(\Theta^r(\mathcal{K}, \nu), L_2(\nu)) \asymp n^{-r}, \quad \nu \in V, \quad r \in [\alpha, \beta]. \tag{4.18}$$

Therefore, by Theorem 1.2

$$t_m(\mathcal{M}(r, \nu), \delta, c) \gg m^{-\frac{2r}{1+2r}}.$$

Choosing $w = 1$ we get from Theorem 4.4 that for any $\rho \in \mathcal{M}(r, \nu) \in \mathbb{M}$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)}^2 \ge C_1(\mathbb{M})(\ln m) t_m(\mathcal{M}(r, \nu), \delta, c)\} \le C_2(\mathbb{M}) m^{-c_1},$$

provided that A is big enough. This indicates that the estimator $f_{\mathbf{z}}^A$ is an universal estimator in a weak sense for the collection $\mathbb{M}$.

5. Big jump estimators. Convex compacts

We will use the following theorem that is a corollary of Theorem 2.3.

Theorem 5.1. *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$. Assume that ρ, $\mathcal{H}$ satisfy (1.5). Then for all $\epsilon > 0$ with probability at least*

$$p(\mathcal{H}, m, \epsilon) := 1 - N(\mathcal{H}, \epsilon/(8M), \mathcal{C}(X)) \exp\Big(-\frac{m\epsilon}{40M^2}\Big) \tag{5.1}$$

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

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f) \le \epsilon.$$

Proof. By Theorem 2.3 (with the parameter 2ϵ) with probability at least $p(\mathcal{H}, m, \epsilon)$ we have

$$\mathcal{E}(f) \le 2\mathcal{E}_{\mathbf{z}}(f) + 2\epsilon + \mathcal{E}(f_{\mathcal{H}}) - 2\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}).$$

Taking into account that for any $f \in \mathcal{H}$ one has $\mathcal{E}(f) \ge \mathcal{E}(f_{\mathcal{H}})$ we obtain

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f) \le \epsilon.$$

The following direct corollary of Theorem 5.1 has been formulated as Corollary 2.1 in Section 2.

Corollary 5.1. *Under assumptions of Theorem 5.1 we have*

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}}) \le \epsilon \tag{5.2}$$

with probability at least $p(\mathcal{H}, m, \epsilon)$.

We present a scheme of building universal estimators based on convex hypothesis spaces. Let $\mathcal{V} := \{V_s\}_{s=0}^{\infty}$ be a sequence of compact convex sets in $\mathcal{C}(X)$. Assume that

$$N(V_s, \epsilon, \mathcal{C}(X)) \le (a2^s(1+1/\epsilon))^{2^s}, \quad V_s \subset aU(\mathcal{C}(X)), \quad s = 0, 1, \dots. \tag{5.3}$$

It will be convenient for us to assume that V_0 consists of only one element f_0. Let ρ be such that all pairs ρ, V_s, $s = 0, 1, \dots$, satisfy (1.5). Then for $\epsilon_s := \frac{A2^s \ln m}{m}$ we get from (5.1) and (5.3) that

$$\sum_{s=0}^{[\log m]} (1 - p(V_s, m, \epsilon_s)) \le m^{-c_1(M)A} \tag{5.4}$$

provided $A \ge A_0(a, M)$.

We take two parameters $A \ge A_0(a, M)$ and K and build an estimator $f_{\mathbf{z}} := f_{\mathbf{z}}(A, K)$ in the following way. Denote

$$\Delta_{\mathbf{z},s} := \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{s-1}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s}).$$

First, if

$$\Delta_{\mathbf{z},s} \le (A+K)\frac{2^s \ln m}{m} + 2A^{1/2}\big(\frac{\ln m}{m}\big)^{1/2}, \quad s = 1, \dots, [\log m] \tag{5.5}$$

then we set $f_{\mathbf{z}} := f_0$.

Second, if (5.5) is not satisfied then we let $l \in [1, \log m]$ be such that for $s \in (l, \log m]$

$$\Delta_{\mathbf{z},s} \le (A+K)\frac{2^s \ln m}{m} + 2A^{1/2}\big(\frac{\ln m}{m}\big)^{1/2}, \tag{5.6}$$

and

$$\Delta_{\mathbf{z},l} > (A+K)\frac{2^l \ln m}{m} + 2A^{1/2}\big(\frac{\ln m}{m}\big)^{1/2}. \tag{5.7}$$

Then we set $f_{\mathbf{z}} := f_{\mathbf{z},V_l}$.

We will prove that this estimator is universal for the following collection of classes. We define a class $W_\nu^r(\mathcal{V}, D)$ as the set of f that satisfy the estimate:

$$d(f, V_s)_{L_2(\nu)} \le D2^{-rs}, \quad s = 0, 1, \dots.$$

We denote

$$\mathcal{W}[\mathcal{V}, D] := \{W_\nu^r(\mathcal{V}, D) : r \le 1/2, \quad \nu \text{ is any Borel measure}\}.$$

Theorem 5.2. *Let $\mathcal{V} = \{V_s\}_{s=0}^{\infty}$ be a sequence of compact convex sets in $\mathcal{C}(X)$ satisfying (5.3). Assume ρ is such that all pairs ρ, V_s, $s = 0, 1, \dots$, satisfy (1.5). For $D > 0$ we set $K := 3D^2$. Then there exists $A_1(a, M)$ such that the estimator $f_{\mathbf{z}} = f_{\mathbf{z}}(A, K)$ with parameters $A \ge A_1(a, M)$, K has the following property. For any $f_\rho \in W^r_{\rho_X}(\mathcal{V}, D)$, $r \le 1/2$ we have*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)} \ge C(D)A\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}} + 4(\log m)\big(\frac{A \ln m}{m}\big)^{1/2}\} \le m^{-c(M)A}.$$

Proof. First of all we use the Bernstein inequality (1.6) and obtain

$$\rho^m\{\mathbf{z} : \max_{0 \le s \le \log m} |\mathcal{E}(f_{V_s}) - \mathcal{E}_{\mathbf{z}}(f_{V_s})| \le \big(\frac{A \ln m}{m}\big)^{1/2}\} \ge 1 - m^{-c_2(M)A} \tag{5.8}$$

provided $A \ge A_0(M)$. We set $A_1(a, M) := \max(A_0(M), A_0(a, M), 1)$.

Second, we use Theorem 2.4 with $\mathcal{H} = V_s$, $\epsilon = A2^s(\ln m)/m$ and obtain

$$\mathcal{E}(f_{\mathbf{z},V_s}) - \mathcal{E}(f_{V_s}) \le A2^s(\ln m)/m, \quad s \in [0, \log m] \tag{5.9}$$

with probability at least $1 - m^{-c_3(M)A}$.

We begin with the case when (5.5) is satisfied and, therefore, $f_{\mathbf{z}} = f_0$. We have

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_0) - \mathcal{E}(f_\rho) = \sum_{s=1}^{\infty}(\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s})). \tag{5.10}$$

Using the assumption $f_\rho \in W^r_{\rho_X}(\mathcal{V}, D)$ we obtain

$$\mathcal{E}(f_{V_n}) - \mathcal{E}(f_\rho) = \|f_{V_n} - f_\rho\|^2_{L_2(\rho_X)} \le D^2 2^{-2rn}, \quad n = 0, 1, \dots$$

and

$$\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s}) \le D^2 2^{-2rs}(1 + 2^{2r}) \le 3D^2 2^{-2rs}, \quad s = 1, \dots. \tag{5.11}$$

We now estimate $\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s})$, $s \in [1, \log m]$, using (5.5). We rewrite

$$\begin{aligned}\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s}) &= (\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}_{\mathbf{z}}(f_{V_{s-1}})) + (\mathcal{E}_{\mathbf{z}}(f_{V_{s-1}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{s-1}})) \\ &+(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{s-1}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s})) + (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s}) - \mathcal{E}_{\mathbf{z}}(f_{V_s})) + (\mathcal{E}_{\mathbf{z}}(f_{V_s}) - \mathcal{E}(f_{V_s})) \\ &=: (e_1) + (e_2) + (\Delta_{\mathbf{z},s}) + (e_3) + (e_4).\end{aligned} \tag{5.12}$$

Let us define Λ to be the set of all $\mathbf{z}$ such that the following relations hold

$$\max_{0 \le s \le \log m} |\mathcal{E}(f_{V_s}) - \mathcal{E}_{\mathbf{z}}(f_{V_s})| \le \big(\frac{A \ln m}{m}\big)^{1/2}; \tag{5.13}$$

$$\mathcal{E}(f_{\mathbf{z},V_s}) - \mathcal{E}(f_{V_s}) \le \frac{A2^s \ln m}{m}, \quad s \in [1, \log m]; \tag{5.14}$$

$$\mathcal{E}_{\mathbf{z}}(f_{V_s}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s}) \le \frac{A2^s \ln m}{m}, \quad s \in [1, \log m]. \tag{5.15}$$

Then by (5.8), (5.9), Corollary 5.1 and (5.4) we get

$$\rho^m(\Lambda) \ge 1 - 3m^{-c_4(M)A}. \tag{5.16}$$

For $\mathbf{z} \in \Lambda$ we have from (5.13)

$$e_1 + e_4 \le 2\big(\frac{A \ln m}{m}\big)^{1/2}, \quad s \in [1, \log m]. \tag{5.17}$$

From (5.15) we obtain

$$e_2 \le \frac{A2^s \ln m}{m}, \quad s \in [1, \log m]. \tag{5.18}$$

From the definition of $f_{\mathbf{z},V_s}$ we have $e_3 \le 0$. We have proved (see (5.11) and a combination of (5.12), (5.17), (5.18)) the following estimate for $\mathbf{z} \in \Lambda$ satisfying (5.5):

$$\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s}) \le \min\{3D^2 2^{-2rs}, (2A+K)\frac{2^s \ln m}{m} + 4\big(\frac{A \ln m}{m}\big)^{1/2}\}, \quad s \in [1, \log m].$$

We use the first estimate for s such that $2^{s(1+2r)} \ge m/\ln m$ and use the second estimate for the remaining s. Summing up these inequalities we get from (5.10) for $r \le 1/2$

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le C(D)A\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}} + 4(\log m)\big(\frac{A \ln m}{m}\big)^{1/2}. \tag{5.19}$$

We now proceed to the case $\mathbf{z} \in \Lambda$ and (5.5) is not satisfied. In this case $f_{\mathbf{z}} = f_{\mathbf{z},V_l}$. We write

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_{\mathbf{z},V_l}) - \mathcal{E}(f_\rho) = \mathcal{E}(f_{\mathbf{z},V_l}) - \mathcal{E}(f_{V_l}) + \sum_{s=l+1}^{\infty} (\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s})). \tag{5.20}$$

The sum $\sum_{s=l+1}^{\infty}(\cdot)$ in the right side of (5.20) can be estimated similar to (5.19) (we use (5.6) in place of (5.5)):

$$\sum_{s=l+1}^{\infty} (\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s})) \le C(D)A\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}} + 4(\log m)\big(\frac{A \ln m}{m}\big)^{1/2}. \tag{5.21}$$

In order to estimate $\mathcal{E}(f_{\mathbf{z},V_l}) - \mathcal{E}(f_{V_l})$ we will obtain an upper estimate for l. On the one hand $\Delta_{\mathbf{z},l}$ satisfies (5.7). On the other hand we have

$$\Delta_{\mathbf{z},l} = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{l-1}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_l}) \tag{5.22}$$

$$= (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{l-1}}) - \mathcal{E}_{\mathbf{z}}(f_{V_{l-1}})) + (\mathcal{E}_{\mathbf{z}}(f_{V_{l-1}}) - \mathcal{E}(f_{V_{l-1}})) + (\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l}))$$

$$+(\mathcal{E}(f_{V_l}) - \mathcal{E}_{\mathbf{z}}(f_{V_l})) + (\mathcal{E}_{\mathbf{z}}(f_{V_l}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_l})) =: (\delta_1) + (\delta_2) + (\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l})) + (\delta_3) + (\delta_4).$$

From the definition of $f_{\mathbf{z},V_{s-1}}$ we get $\delta_1 \le 0$. By (5.13) we have for $\mathbf{z} \in \Lambda$

$$\delta_2 + \delta_3 \le 2\big(\frac{A \ln m}{m}\big)^{1/2}. \tag{5.23}$$

By (5.11) we have

$$\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l}) \le K 2^{-2rl}. \tag{5.24}$$

By (5.15) we obtain for $\mathbf{z} \in \Lambda$

$$\delta_4 \le \frac{A 2^l \ln m}{m}. \tag{5.25}$$

Combining (5.22)–(5.25) we get

$$\Delta_{\mathbf{z},l} \le 2\big(\frac{A \ln m}{m}\big)^{1/2} + K 2^{-2rl} + \frac{A 2^l \ln m}{m}.$$

Comparing this with (5.7) we conclude that

$$2^{-2rl} \ge 2^l (\ln m)/m. \tag{5.26}$$

By (5.14) we have

$$\mathcal{E}(f_{\mathbf{z},V_l}) - \mathcal{E}(f_{V_l}) \le A 2^l (\ln m)/m. \tag{5.27}$$

Substituting (5.26) into (5.27) we get from (5.21) and (5.27) that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le C(D) A \big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}} + 4(\log m)\big(\frac{A \ln m}{m}\big)^{1/2}. \quad \square$$

6. Big jump estimators. Nonconvex compacts

We now present a general scheme of building universal estimators. Let $\mathcal{V} := \{V_s\}_{s=0}^{\infty}$ be a sequence of compact sets in $\mathcal{C}(X)$. Assume that

$$N(V_s, \epsilon, \mathcal{C}(X)) \le (a2^{bs}(1+1/\epsilon))^{2^s}, \quad V_s \subset aU(\mathcal{C}(X)), \quad s = 0, 1, \dots. \tag{6.1}$$

It will be convenient for us to assume that V_0 consists of only one element f_0. Let ρ be such that all pairs ρ, V_s, $s = 0, 1, \dots$, satisfy (1.5). We take two parameters $A \ge A_0(a, M)$ and K and define an estimator $f_{\mathbf{z}}$ in the same way as in Section 5 (see (5.5)–(5.7)). We will prove that this estimator is universal for the collection $\mathcal{W}[\mathcal{V}, D]$ of classes defined in Section 5.

Theorem 6.1. *Let $\mathcal{V} = \{V_s\}_{s=0}^{\infty}$ be a sequence of compact sets in $\mathcal{C}(X)$ satisfying (6.1). Assume ρ is such that all pairs ρ, V_s, $s = 0, 1, \dots$, satisfy (1.5). For $D > 0$ we set $K := 3D^2$. Then there exists $A_1(a, b, M)$ such that the estimator $f_{\mathbf{z}} = f_{\mathbf{z}}(A, K)$ with parameters $A \ge A_1(a, b, M)$, K has the following property. For any $f_\rho \in W^r_{\rho_X}(\mathcal{V}, D)$, $r \le 1/2$ we have*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|^2_{L_2(\rho_X)} \ge C(D)A(\log m)\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}}\} \le m^{-c(M)A}.$$

Proof. Let s_r be the minimum s satisfying $2^s(\ln m)/m \ge 2^{-2rs}$. We define Λ to be the set of all $\mathbf{z}$ such that the following relations hold

$$\max_{0 \le s \le \log m} |\mathcal{E}(f_{V_s}) - \mathcal{E}_{\mathbf{z}}(f_{V_s})| \le \big(\frac{A \ln m}{m}\big)^{1/2}; \tag{6.1}$$

$$\mathcal{E}(f_{\mathbf{z},V_s}) - \mathcal{E}(f_{V_s}) \le \frac{A2^{\max(s,s_r)} \ln m}{m}, \quad s \in [1, \log m]; \tag{6.2}$$

$$\mathcal{E}_{\mathbf{z}}(f_{V_s}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s}) \le \frac{A2^{\max(s,s_r)} \ln m}{m}, \quad s \in [1, \log m]. \tag{6.3}$$

By Bernstein's inequality (6.1) holds with probability $\ge 1 - m^{-c_1(M)A}$ provided $A \ge A_0(M)$. By Theorem 3.6 (6.2) holds with probability $\ge 1 - m^{-c_1(M)A}$. By Corollary 3.2 (6.3) holds with probability $\ge 1 - m^{-c_1(M)A}$. Therefore,

$$\rho^m(\Lambda) \ge 1 - 3m^{-c_1(M)A}$$

provided $A \ge C_1(a, b, M)$.

We begin with the case when (5.5) is satisfied and, therefore, $f_{\mathbf{z}} = f_0$. Similar to (5.11) we have

$$\mathcal{E}(f_{V_{s-1}}) - \mathcal{E}(f_{V_s}) \le D^2 2^{-2rs}(1 + 2^{2r}) \le 3D^2 2^{-2rs}, \quad s = 1, \dots. \tag{6.4}$$

We now estimate $\mathcal{E}(f_{V_{s-1}})-\mathcal{E}(f_{V_s})$, $s\in[1,\log m]$, using (5.5). Similar to (5.12) we rewrite

$$\mathcal{E}(f_{V_{s-1}})-\mathcal{E}(f_{V_s})=(\mathcal{E}(f_{V_{s-1}})-\mathcal{E}_{\mathbf{z}}(f_{V_{s-1}}))+(\mathcal{E}_{\mathbf{z}}(f_{V_{s-1}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{s-1}}))$$
$$+(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{s-1}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s}))+(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_s})-\mathcal{E}_{\mathbf{z}}(f_{V_s}))+(\mathcal{E}_{\mathbf{z}}(f_{V_s})-\mathcal{E}(f_{V_s}))$$
$$=:(e_1)+(e_2)+(\Delta_{\mathbf{z},s})+(e_3)+(e_4). \tag{6.5}$$

From (6.1) we have for $\mathbf{z}\in\Lambda$

$$e_1+e_4\le 2\big(\frac{A\ln m}{m}\big)^{1/2},\quad s\in[1,\log m]. \tag{6.6}$$

By (6.3) we obtain for $\mathbf{z}\in\Lambda$

$$e_2\le\frac{A2^{s_r}\ln m}{m},\quad s\in[1,s_r]. \tag{6.7}$$

Substituting (6.6), (6.7) into (6.5) we obtain the following estimate for $\mathbf{z}\in\Lambda$ satisfying (5.5):

$$\mathcal{E}(f_{V_{s-1}})-\mathcal{E}(f_{V_s})\le(2A+K)\frac{2^{s_r}\ln m}{m}+4\big(\frac{A\ln m}{m}\big)^{1/2},\quad s\in[1,s_r]. \tag{6.8}$$

We use (6.4) for $s\ge s_r$ and use (6.8) for the remaining s. Summing up these inequalities we get from (5.10) for $r\le 1/2$

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)\le C(D)A(\log m)\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}}. \tag{6.9}$$

We now proceed to the case $\mathbf{z}\in\Lambda$ and (5.5) is not satisfied. In this case $f_{\mathbf{z}}=f_{\mathbf{z},V_l}$. We write

$$\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(f_\rho)=\mathcal{E}(f_{\mathbf{z},V_l})-\mathcal{E}(f_\rho)=\mathcal{E}(f_{\mathbf{z},V_l})-\mathcal{E}(f_{V_l})+\sum_{s=l+1}^{\infty}(\mathcal{E}(f_{V_{s-1}})-\mathcal{E}(f_{V_s})). \tag{6.10}$$

The sum $\sum_{s=l+1}^{\infty}(\cdot)$ in the right side of (6.10) can be estimated similar to (6.9):

$$\sum_{s=l+1}^{\infty}(\mathcal{E}(f_{V_{s-1}})-\mathcal{E}(f_{V_s}))\le C(D)A(\log m)\big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}}. \tag{6.11}$$

In order to estimate $\mathcal{E}(f_{\mathbf{z},V_l})-\mathcal{E}(f_{V_l})$ we need an upper estimate for l. Suppose $l\le s_r$. Then by (6.2) we obtain for $\mathbf{z}\in\Lambda$

$$\mathcal{E}(f_{\mathbf{z},V_l})-\mathcal{E}(f_{V_l})\le\frac{A2^{s_r}\ln m}{m}.$$

Now, suppose $l \ge s_r$. On the one hand $\Delta_{\mathbf{z},l}$ satisfies (4.7). On the other hand we have

$$\Delta_{\mathbf{z},l} = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{l-1}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_l}) \tag{6.12}$$

$$= (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_{l-1}}) - \mathcal{E}_{\mathbf{z}}(f_{V_{l-1}})) + (\mathcal{E}_{\mathbf{z}}(f_{V_{l-1}}) - \mathcal{E}(f_{V_{l-1}})) + (\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l}))$$

$$+(\mathcal{E}(f_{V_l}) - \mathcal{E}_{\mathbf{z}}(f_{V_l})) + (\mathcal{E}_{\mathbf{z}}(f_{V_l}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},V_l})) =: (\delta_1) + (\delta_2) + (\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l})) + (\delta_3) + (\delta_4).$$

From the definition of $f_{\mathbf{z},V_{s-1}}$ we get $\delta_1 \le 0$. By (6.1) we have for $\mathbf{z} \in \Lambda$

$$\delta_2 + \delta_3 \le 2\big(\frac{A \ln m}{m}\big)^{1/2}. \tag{6.13}$$

By (6.4) we have

$$\mathcal{E}(f_{V_{l-1}}) - \mathcal{E}(f_{V_l}) \le K 2^{-2rl}. \tag{6.14}$$

By (6.3) we obtain for $\mathbf{z} \in \Lambda$

$$\delta_4 \le \frac{A 2^l \ln m}{m}. \tag{6.15}$$

Combining (6.12)–(6.15) we get

$$\Delta_{\mathbf{z},l} \le 2\big(\frac{A \ln m}{m}\big)^{1/2} + K 2^{-2rl} + \frac{A 2^l \ln m}{m}.$$

Comparing this with (5.7) we conclude that

$$2^{-2rl} \ge 2^l (\ln m)/m. \tag{6.16}$$

By (6.2) we have for $\mathbf{z} \in \Lambda$

$$\mathcal{E}(f_{\mathbf{z},V_l}) - \mathcal{E}(f_{V_l}) \le 2^l (\ln m)/m. \tag{6.17}$$

Substituting (6.16) into (6.17) we get from (6.17) and (6.11) that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le C(D) A (\log m) \big(\frac{\ln m}{m}\big)^{\frac{2r}{1+2r}}. \quad \square$$

7. Some examples

In Sections 4–6 we presented two different ways of construction of universal estimators: the penalized least squares estimators and the big jump estimators. Both these methods are based on a given sequence of compacts in $\mathcal{C}(X)$. In Section 4 we considered a collection $\mathcal{K}(a,b)$ of compacts K_n in $\mathcal{C}(X)$ satisfying

$$N(K_n,\epsilon,\mathcal{C}(X)) \le (an^b/\epsilon)^n, \quad n=1,2,\dots. \tag{7.1}$$

In Section 6 we used a collection $\mathcal{V} := \mathcal{V}(a,b) := \{V_s\}_{s=0}^{\infty}$ of compacts V_s in $\mathcal{C}(X)$ such that

$$N(V_s,\epsilon,\mathcal{C}(X)) \le (a2^{bs}/\epsilon)^{2^s}, \quad s=0,1,\dots. \tag{7.2}$$

It is clear that the sequence $\mathcal{V}(a,b)$ can be seen as a dyadic subsequence of $\mathcal{K}(a,b)$. In Section 5 we considered a particular case $\mathcal{V}(a,1)$. Therefore, in the discussion that follows we will describe different examples of sequences $\mathcal{K}(a,b)$.

We begin with a construction based on the concept of the Kolmogorov width. This construction has been used in [DKPT1,2]. Kolmogorov's n-width for a centrally symmetric compact set F in a Banach space B is defined as follows

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

where $\inf_L$ is taken over all n-dimensional subspaces of B. In other words the Kolmogorov n-width gives the best possible error in approximating a compact set F by n-dimensional linear subspaces.

Example 1. Let $\mathcal{L} = \{L_n\}_{n=1}^{\infty}$ be a sequence of n-dimensional subspaces of $\mathcal{C}(X)$. For $Q>0$ we define

$$K_n := QU(\mathcal{C}(X)) \cap L_n = \{f\in L_n : \|f\|_{\mathcal{C}(X)} \le Q\}, \quad n=1,2,\dots.$$

Then it is well known [P] that

$$N(K_n,\epsilon,\mathcal{C}(X)) \le Q^n(1+2/\epsilon)^n \le (2Q(1+1/\epsilon))^n.$$

We note that $\{K_n\}_{n=1}^{\infty} = \mathcal{K}(2Q,0)$. Therefore, Theorem 4.4 applies to this sequence of compacts. Let us discuss the condition

$$d(f_\rho,K_n)_{L_2(\rho_X)} \le A^{1/2}n^{-r}, \quad n=1,2,\dots, \tag{7.3}$$

from Theorem 4.4. We compare (7.3) with a standard in approximation theory condition

$$d(f_\rho,L_n)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n=1,2,\dots, \quad f_\rho \in DU(\mathcal{C}(X)). \tag{7.4}$$

First of all we observe that (7.4) implies that there exists $\varphi_n \in L_n$, $\|\varphi_n\|_{\mathcal{C}(X)} \le 2D$, such that

$$\|f_\rho - \varphi_n\|_{\mathcal{C}(X)} \le Dn^{-r}.$$

Thus (7.4) implies

$$d(f_\rho,K_n)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n=1,2,\dots \tag{7.5}$$

provided $Q \ge 2D$. Also, (7.5) implies (7.3) provided $A^{1/2} \ge D$. Therefore, Theorem 4.4 can be used for f_ρ satisfying (7.4). We formulate this result as a theorem.

Theorem 7.1. *Let $\mathcal{L} = \{L_n\}_{n=1}^{\infty}$ be a sequence of n-dimensional subspaces of $\mathcal{C}(X)$. For given positive numbers D, M_1, $M := M_1 + D$ there exists $A_0 := A_0(D, M)$ with the following property. For any $A \ge A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \le M_1$ a.e. with respect to ρ and*

$$d(f_\rho, L_n)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n = 1, 2, \dots, \quad f_\rho \in DU(\mathcal{C}(X))$$

we have for $\eta \ge A^{1/2}\left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)} \ge 4A^{1/2}\eta\} \le Ce^{-c(M)m\eta^2}. \tag{7.6}$$

Theorem 7.1 is an extension of Theorem 4.10 from [DKPT2]. Theorem 4.10 from [DKPT2] gives (7.6) with $e^{-c(M)m\eta^2}$ replaced by $e^{-c(M)m\eta^4}$ under an extra restriction $r \le 1/2$.

Example 2. In the previous example we worked in the $\mathcal{C}(X)$ space. We now want to replace (7.4) by a weaker condition

$$d(f_\rho, L_n)_{L_2(\rho_X)} \le Dn^{-r}, \quad n = 1, 2, \dots, \quad f_\rho \in DU(L_2(\rho_X)). \tag{7.7}$$

This condition is compatible with the condition (7.3) (from Theorem 4.4) in the sense of approximation in the $L_2(\rho_X)$ norm. However, the conditions (7.7) and (7.3) differ in the sense of the approximation set: it is a linear subspace L_n in (7.7) and a compact subset of $\mathcal{C}(X)$ in (7.3). In Example 1 approximation (7.4) by a linear subspace automatically provided approximation (7.3) by a suitable compact of $\mathcal{C}(X)$. It is clear that similarly to Example 1 approximation (7.7) by a linear subspace L_n provides approximation (7.3) by a compact $K_n \subset L_n$ of the $L_2(\rho_X)$ instead of the $\mathcal{C}(X)$. We cannot apply Theorem 4.4 in such a situation. In order to overcome this difficulty we impose an extra restrictions on the sequence $\mathcal{L}$ and on the measure ρ. We discuss the setting from [KT2]. Let $\mathcal{B}(X)$ be a Banach space with the norm $\|f\|_{\mathcal{B}(X)} := \sup_{x \in X} |f(x)|$. Let $\{L_n\}_{n=1}^{\infty}$ be a given sequence of n-dimensional linear subspaces of $\mathcal{B}(X)$ such that L_n is also a subspace of each $L_\infty(\mu)$, where μ is a Borel probability measure on X, $n = 1, 2, \dots$. Assume that n-dimensional linear subspaces L_n have the following property: for any Borel probability measure μ on X one has

$$\|P_{L_n}^\mu\|_{\mathcal{B}(X) \to \mathcal{B}(X)} \le K, \quad n = 1, 2, \dots, \tag{7.8}$$

where P_L^μ is the operator of $L_2(\mu)$ projection onto L. Then our standard assumption $|y| \le M_1$ implies $\|f_\rho\|_{L_\infty(\rho_X)} \le M_1$ and (7.7), (7.8) give

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

where

$$K_n := (K+1)M_1 U(\mathcal{B}(X)) \cap L_n.$$

We note that Theorem 4.4 holds for compacts satisfying (3.12) in the $\mathcal{B}(X)$ norm instead of $\mathcal{C}(X)$ norm. Thus, as a corollary of Theorem 4.4 we obtain the following result.

Theorem 7.2. *Let $\mathcal{L} = \{L_n\}_{n=1}^{\infty}$ be a sequence of n-dimensional subspaces of $\mathcal{B}(X)$ satisfying (7.8). For given positive numbers D, M_1, $M := M_1 + D$ there exists $A_0 := A_0(K, D, M)$ with the following property. For any $A \ge A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \le M_1$ a.e. with respect to ρ and*

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

we have for $\eta \ge \eta_m := A^{1/2}\left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)} \ge 4A^{1/2}\eta\} \le Ce^{-c(M)m\eta^2}. \tag{7.9}$$

Theorem 7.2 is an extension of Theorem 4.3 from [KT2]. Theorem 4.3 from [KT2] gives (7.9) with $4A^{1/2}\eta$ replaced by $C(D)\eta_m$ and $e^{-c(M)m\eta^2}$ replaced by $m^{-c(M)A}$ under an extra restriction $r \le 1/2$.

Remark 7.1. *In Theorem 7.2 we can replace the assumption that $\mathcal{L}$ satisfies (7.8) for all Borel probability measures μ by the assumption that (7.8) is satisfied for $\mu \in \mathcal{M}$ and add the assumption $\rho_X \in \mathcal{M}$.*

Example 3. Our construction here is based on the concept of nonlinear Kolmogorov's (N, n)-width ([T1]):

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

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

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

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

Let $\mathbb{L} := \{\mathcal{L}_n\}_{n=1}^{\infty}$ be a sequence of collections $\mathcal{L}_n := \{L_n^j\}_{j=1}^{N_n}$ of n-dimensional subspaces L_n^j of $\mathcal{C}(X)$. Assume $N_n \le n^{bn}$. For $Q > 0$ we now consider

$$K_n := \bigcup_{j=1}^{N_n} (QU(\mathcal{C}(X)) \cap L_n^j).$$

Then $\{K_n\}_{n=1}^{\infty} = \mathcal{K}(2Q, b)$. It is also clear that the condition

$$\min_{1 \le j \le N_n} d(f_\rho, L_n^j)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n = 1, 2, \dots, \quad f_\rho \in DU(\mathcal{C}(X)) \tag{7.10}$$

implies

$$d(f_\rho, K_n)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n = 1, 2, \dots,$$

provided $Q \ge 2D$.

We have the following analogue of Theorem 7.1.

Theorem 7.3. *Let $\mathbb{L} := \{\mathcal{L}_n\}_{n=1}^{\infty}$ be a sequence of collections $\mathcal{L}_n := \{L_n^j\}_{j=1}^{N_n}$ of n-dimensional subspaces L_n^j of $\mathcal{C}(X)$. Assume $N_n \le n^{bn}$. For given positive numbers D, M_1, $M := M_1 + D$ there exists $A_0 := A_0(b, D, M)$ with the following property. For any $A \ge A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \le M_1$ a.e. with respect to ρ and*

$$\min_{1\le j\le N_n} d(f_\rho, L_n^j)_{\mathcal{C}(X)} \le Dn^{-r}, \quad n = 1, 2, \dots, \quad f_\rho \in DU(\mathcal{C}(X))$$

we have for $\eta \ge A^{1/2}\left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

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

Example 4. In this example we apply the ideas of Examples 2 and 3 for nonlinear m-term approximation with regard to a given countable dictionary. Let $\Psi := \{\psi_n\}_{n=1}^{\infty}$ be a system of functions ψ_n from $\mathcal{B}(X)$. Let $\gamma \ge 0$ and let $\mathcal{M}(\gamma)$ be a set of Borel probability measures μ such that all ψ_n are μ-measurable and

$$\|\sum_{n=1}^{N} a_n\psi_n\|_{\mathcal{B}(X)} \le C_1N^{\gamma}\|\sum_{n=1}^{N} a_n\psi_n\|_{L_2(\mu)}. \tag{7.11}$$

We fix a parameter $q \ge 1$ and define the best m-term approximation with depth m^q as follows

$$\sigma_{m,q}(f, \Psi)_{L_2(\mu)} := \inf_{c_i; n_i \le m^q} \|f - \sum_{i=1}^{m} c_i\psi_{n_i}\|_{L_2(\mu)}.$$

For a fixed $Q > 0$ that will be chosen later we now consider

$$K_n(Q) := \{f : f = \sum_{i=1}^{n} a_i\psi_{n_i}, \quad n_i \le n^q, \quad i = 1, \dots, n, \quad \|f\|_{\mathcal{B}(X)} \le Q\}.$$

Then

$$N(K_n(Q), \epsilon, \mathcal{B}(X)) \le (Q(1 + 2/\epsilon))^n \binom{n^q}{n} \le (2Qn^q(1 + 1/\epsilon))^n. \tag{7.12}$$

Suppose we have for $\mu \in \mathcal{M}(\gamma)$

$$\sigma_{m,q}(f, \Psi)_{L_2(\mu)} \le Dn^{-r}, \quad f \in DU(L_2(\mu)).$$

Then there exists φ_n of the form

$$\varphi_n = \sum_{i=1}^{n} a_i\psi_{n_i}, \quad n_i \le n^q, \quad i = 1, \dots, n, \quad \|\varphi_n\|_{L_2(\mu)} \le 2D$$

such that

$$\|f - \varphi_n\|_{L_2(\mu)} \le Dn^{-r}.$$

Next, by our assumption (7.11) we get

$$\|\varphi_n\|_{\mathcal{B}(X)} \le 2DC_1n^{\gamma q}.$$

Therefore $\varphi_n \in K_n(2DC_1n^{\gamma q})$. The inequality (7.12) implies that

$$\{K_n(2DC_1n^{\gamma q})\} = \mathcal{K}(4DC_1, (1 + \gamma)q).$$

Consequently, Theorem 4.4 applies in this situation. We formulate the result as a theorem.

Theorem 7.4. *Let Ψ and $\mathcal{M}(\gamma)$ be as above. For given positive numbers q, D, M_1, $M := M_1 + D$, there exists $A_0 := A_0(\gamma, C_1, q, D, M)$ with the following property. For any $A \ge A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \le M_1$ a.e. with respect to ρ, $\rho_X \in \mathcal{M}(\gamma)$, and*

$$\sigma_{n,q}(f_\rho, \Psi)_{L_2(\rho_X)} \le Dn^{-r}, \quad n = 1, 2, \dots, \quad f_\rho \in DU(\mathcal{B}(X))$$

we have for $\eta \ge A^{1/2}\left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

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

Remark 7.2. *In Theorem 7.4 the condition (7.11) can be replaced by the following weaker condition. Let $m_i \in [1, n^q]$, $i = 1, \dots, n$, and*

$$L_n := \operatorname{span}\{\psi_{m_i}\}_{i=1}^n.$$

Assume that for $\mu \in \mathcal{M}(\gamma)$

$$\|P_{L_n}^\mu\|_{\mathcal{B}(X)\to\mathcal{B}(X)} \le C_1 n^{qn}.$$

References

[B] A.R. Barron, *Complexity regularizasion with applications to artificial neural networks*, In Nonparametric Functional Estimation (G. Roussas, ed.), Kluwer, Dordrecht (1991), 561–576.

[BBM] A. Barron, L. Birgé, P. Massart, *Risk bounds for model selection via penalization*, Probability Theory and Related Fields **113** (1999), 301–413.

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

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

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

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

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

[EPP] T. Evgeniou, M. Pontil and T. Poggio, *Regularization Networks and Support Vector Machines*, Advances in Comput. Math. **13** (2000), 1–50.

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

[KT1] S. Konyagin and V. Temlyakov, *Some error estimates in Learning Theory*, IMI Preprints **05** (2004), 1–18.

[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, Viena (2002), 5–62.

[LBM] W.-S. Lee, P. Bartlett, and R. Williamson, *The importance of convexity in learning with square loss*, IEE Transactions on Information Theory **44** (1998), 1974–1980.

[P] G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.

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

[T1] V.N. Temlyakov, *Nonlinear Kolmogorov's widths*, Matem. Zametki **63** (1998), 891–902.

[T2] V.N. Temlyakov, *Optimal Estimators in Learning Theory*, IMI Preprints **23** (2004), 1–29.

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

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