2004:09

The entropy in the learning theory.
Error estimates

S.V. Konyagin and V.N. Temlyakov
THE ENTROPY IN THE LEARNING THEORY. ERROR ESTIMATES

S.V. KONYAGIN AND V.N. TEMLYAKOV

Abstract. We continue investigation of some problems in learning theory in the setting formulated by F. Cucker and S. Smale [CS]. The goal is to find an estimator \( f_z \) on the base of given data \( z := ((x_1, y_1), \ldots, (x_m, y_m)) \) that approximates well the regression function \( f_\rho \) of an unknown Borel probability measure \( \rho \) defined on \( Z = X \times Y \). We assume that \( f_\rho \) belongs to a function class \( W \). It is known from the previous works that the behavior of the entropy numbers \( \epsilon_n(W, C) \) of \( W \) in the uniform norm \( C \) plays an important role in the above problem.

The standard way of measuring the error between a target function \( f_\rho \) and an estimator \( f_z \) is to use the \( L^2(\rho_X) \) norm (\( \rho_X \) is the marginal probability measure on \( X \) generated by \( \rho \)). This way has been used in the previous papers. We also follow this way in the paper. The use of the \( L^2(\rho_X) \) norm in measuring the error has motivated us to study the case when we make an assumption on the entropy numbers \( \epsilon_n(W, L^2(\rho_X)) \) of \( W \) in the \( L^2(\rho_X) \) norm. This is the main new ingredient of the paper. We construct good estimators in different settings: 1. we know both \( W \) and \( \rho_X \); 2. we know \( W \) and we do not know \( \rho_X \); 3. we only know that \( W \) is from a known collection of classes and we do not know \( \rho_X \). An estimator from the third setting is called universal estimator [DKPT].

1. Introduction

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

Let \( X \subset \mathbb{R}^d, Y \subset \mathbb{R} \) be Borel sets, \( \rho \) be a Borel probability measure on \( Z = X \times Y \). For \( f : X \to Y \) define the error

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

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

\[
f_\rho(x) := \int_Y yd\rho(y|x).
\]
The function $f_\rho$ is known in statistics as the regression function of $\rho$. It is clear that if $f_\rho \in L_2(\rho_X)$ then it minimizes the error $\mathcal{E}(f)$ over all $f \in L_2(\rho_X)$: $\mathcal{E}(f_\rho) \leq \mathcal{E}(f)$, $f \in L_2(\rho_X)$. Thus, in the sense of error $\mathcal{E}(\cdot)$ the regression function $f_\rho$ is the best to describe the relation between inputs $x \in X$ and outputs $y \in Y$. Now, our goal is to find an estimator $f_z$, on the base of given data $z = ((x_1, y_1), \ldots, (x_m, y_m))$ that approximates $f_\rho$ well with high probability. We assume that $(x_i, y_i)$, $i = 1, \ldots, m$ are independent and distributed according to $\rho$. 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 [DKPT] and [KT]. 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], [KT]) is based on the idea of studying approximation of a projection $f_W$ of $f_\rho$ onto $W$. In this case we do not assume that the regression function $f_\rho$ comes from a specific (say, smoothness) class of functions. The second way ([CS], [PS], [DKPT], [KT]) is based on the assumption $f_\rho \in W$. For instance, we may assume that $f_\rho$ has some smoothness. The next step is to find a method for constructing an estimator $f_z$ that provides a good (optimal, near optimal in a certain sense) error $\|f_\rho - f_z\|$ for all $f_\rho \in W$ with high probability with respect to $\rho$. A problem of optimization is naturally broken into two parts: upper estimates and lower estimates. In order to prove upper estimates we need to decide what should be the form of an estimator $f_z$. In other words we need to specify the hypothesis space $\mathcal{H}$ (see [CS], [PS], [KT]) (approximation space [DKPT], [KT]) where an estimator $f_z$ comes from.

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

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

where

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

is the empirical error (risk) of $f$. This $f_{z,\mathcal{H}}$ is called the empirical optimum.

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

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

where $U(B)$ is the unit ball of Banach space $B$. We denote $N(W, \epsilon, B)$ the covering number that is the minimal number of balls of radius $\epsilon$ needed for covering $W$. In the papers [CS], [DKPT], [KT] in the most cases the space $\mathcal{C} := \mathcal{C}(X)$ of continuous functions on a compact $X \subset \mathbb{R}^d$ has been taken as a Banach space $B$. This allowed to formulate all results with assumptions on $W$ independent of $\rho$. In this paper we obtain some results for $B = L_2(\rho_X)$. On the one hand we weaken assumptions on the class $W$ and on the other hand this results in the use of $\rho_X$ in the construction of an estimator. Thus, we have a tradeoff between treating
wider classes and building estimators that are independent of $\rho_X$. We show in Section 4 that in some special cases of interest in applications we can construct universal estimators for wider classes. In [DKPT], [KT] the restrictions on a class $W$ have been imposed in the following form:

\[(1.1) \quad \epsilon_n(W, \mathcal{C}) \leq Dn^{-r}, \quad n = 1, 2, \ldots, \quad W \subset DU(\mathcal{C}).\]

In this paper we impose a weaker restriction

\[(1.2) \quad \epsilon_n(W, L_2(\rho_X)) \leq Dn^{-r}, \quad n = 1, 2, \ldots, \quad W \subset DU(L_2(\rho_X)).\]

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

\[(1.3) \quad E(f) - E(f_\rho) = \| f - f_\rho \|_{L_2(\rho_X)}^2.\]

Thus the choice $\| \cdot \| = \| \cdot \|_{L_2(\rho_X)}$ seems natural. This norm has also been used in [DKPT], [KT] for measuring the error. The use of the $L_2(\rho_X)$ norm in measuring the error is the main reason for us to consider restrictions (1.2) instead of (1.1).

One of important questions discussed in [CS], [DKPT], [KT] is to estimate the defect function $L_z(f) := E(f) - E_z(f)$ of $f \in W$. If $\xi$ is a random variable (a real valued function on a probability space $Z$) then denote

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

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

\[(1.4) \quad \text{Prob}_{\xi \in Z^m}\{ \frac{1}{m} \sum_{i=1}^m \xi(z_i) - E(\xi) \geq \epsilon \} \leq 2 \exp\left(-\frac{m\epsilon^2}{2(\sigma^2(\xi) + M\epsilon/3)}\right).\]

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

$$\text{Prob}_{\xi \in Z^m}\{ |L_z(f)| \leq \epsilon \} \geq 1 - 2 \exp\left(-\frac{m\epsilon^2}{2(\sigma^2 + M^2\epsilon/3)}\right),$$

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

We will assume that $\rho$ and $W$ satisfy the following condition.

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

The following useful inequality has been obtained in [CS].
Theorem 1.2 [CS]. Let $W$ be a compact subset of $\mathcal{C}(X)$. Assume that $\rho$, $W$ satisfy (1.5). Then, for all $\epsilon > 0$

\begin{equation}
\text{Prob}_{z \in \mathbb{Z}^m} \{ \sup_{f \in W} |L_z(f)| \geq \epsilon \} \leq N(W, \epsilon/(8M), \mathcal{C}) 2 \exp \left( \frac{-mc^2}{2(2\sigma^2 + M^2\epsilon/3)} \right).
\end{equation}

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

This theorem contains a factor $N(W, \epsilon/(8M), \mathcal{C})$ that may grow exponentially for classes $W$ satisfying (1.1): $N(W, \epsilon, \mathcal{C}) \leq 2^{(D/\epsilon)^{1/r} + 1}$. A stronger (in a certain sense) estimate than (1.6) has been obtained in [KT] under assumption that $W$ satisfies (1.1).

Theorem 1.3 [KT]. Assume that $\rho$, $W$ satisfy (1.5) and $W$ is such that

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

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

\begin{equation}
\text{Prob}_{z \in \mathbb{Z}^m} \{ \sup_{f \in W} |L_z(f)| \geq \eta \} \leq C(M, \epsilon(W)) \exp(-c(M)m\eta^2)
\end{equation}

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

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

In Section 2 we prove that it is impossible to have even a weaker analog of Theorem 1.3 if we use the $L_2(\rho_X)$ norm instead of the uniform norm $\mathcal{C}$. However, it turned out that we can prove an $L_2(\rho_X)$ analog of Theorem 1.3 for the $\delta$-net $\mathcal{N}_\delta(W)$ of $W$ in the $L_2(\rho_X)$ norm instead of $W$ for $\delta^2 \geq \eta$ (see Theorem 2.2).

It is well known ([CS], [DKPT], [KT]) how estimates of the defect function $L_z(f)$, $f \in \mathcal{H}$, can be used for estimating the error $\mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_\rho)$, $f_\rho \in W$. We prove in Section 2 the following theorem.

Theorem 1.4. Let $f_\rho \in W$ and let $\rho$, $W$ satisfy (1.5) and (1.2) with $r > 1/2$. Then there exists an estimator $f_z$ such that for $A \geq 2$

\begin{equation}
\text{Prob}_{z \in \mathbb{Z}^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3A^{1/2}(\ln(m/m)^{1/2}) \} \geq 1 - C(M, D, r)m^{-c(M)A}.
\end{equation}

Also

\begin{equation}
\text{Prob}_{z \in \mathbb{Z}^m} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_\rho)| \leq 4A^{1/2}(\ln(m/m)^{1/2}) \} \geq 1 - C(M, D, r)m^{-c(M)A}.
\end{equation}

It is interesting to compare this result with the known result from [KT] when we assume (1.1) instead of (1.2).
Theorem 1.5 [KT]. Let \( f_\rho \in W \) and let \( \rho \) and \( W \) satisfy (1.1) and (1.5). Then there exists an estimator \( f_z \) such that for \( A \geq A_0(M, D, r) \)

\[
(1.8) \quad \mathbb{P}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq A m^{-\frac{2r}{1+2r}} \} \geq 1 - \exp(-c(M)A m^{\frac{1}{1+2r}}).
\]

We see that for \( r > 1/2 \) close to 1/2 the exponent 1/2 from (1.7) is close to the exponent \( \frac{2r}{1+2r} \) from (1.8). However, for big \( r \) (1.8) provides much better error estimates than (1.7). We do not know if (1.7) can be improved in this case. Surprisingly, in the case \( r \in (0, 1/2] \) we obtain the error estimates only slightly worse than (1.8) under a weaker assumption (1.2). We prove in Section 3 the following estimates.

Theorem 1.6. Let \( f_\rho \in W \) and let \( \rho, W \) satisfy (1.5) and (1.2). Then there exists an estimator \( f_z \) such that for \( A \geq A_0(M, D, r) \geq 2 \)

\[
\mathbb{P}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3A((\ln m)^3/m)^{1/2} \} \geq 1 - C(M, D)m^{-c(M, D)A^2},
\]

\[
\mathbb{P}_{z \in Z^m} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_\rho)| \leq 4A((\ln m)^3/m)^{1/2} \} \geq 1 - C(M, D)m^{-c(M, D)A^2},
\]

provided \( r = 1/2 \),

\[
\mathbb{P}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3A(\ln m/m)^{\frac{2r}{1+2r}} \} \geq 1 - C(M, D, r)m^{-c(M, D, r)A^{1+\frac{1}{r}}},
\]

\[
\mathbb{P}_{z \in Z^m} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_\rho)| \leq 4A(\ln m/m)^{\frac{2r}{1+2r}} \} \geq 1 - C(M, D, r)m^{-c(M, D, r)A^{1+\frac{1}{r}}},
\]

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

We note that the estimator \( f_z \) from Theorem 1.6 is \( f_z, \mathcal{H} \) with \( \mathcal{H} : = \mathcal{N}_\delta(W, C) \) of \( W \) in the \( L_2(\rho_X) \) norm. The parameter \( \delta(m, r) \) depends on \( m \) and \( r \) that comes from (1.2). Thus, in order to build \( f_z \) from Theorem 1.6 we need to know the class \( W \) (in particular, a parameter \( r \) from (1.2)) and the measure \( \rho_X \). It is clear that if \( W \) satisfies (1.1) then a minimal \( \delta \)-net \( \mathcal{N}_\delta(W, C) \) of \( W \) in the \( C \) norm may serve as a \( \delta \)-net of \( W \) in the \( L_2(\rho_X) \) norm for all \( \rho_X \). Therefore, it is natural (see Theorem 1.5) that if \( W \) satisfies (1.1) then a good estimator \( f_z \) does not depend on \( \rho_X \). In Section 4 we present a special example of interest in applications where we build an estimator \( f_z \) independent of \( \rho_X \) that provides good error estimates for classes \( W \) satisfying approximation properties imposed in the \( L_2(\rho_X) \) norm. The above mentioned example is based on the idea used in [DKPT] of imposing restrictions on the class \( W \) in terms of approximation by linear subspaces rather than in terms of approximation by finite nets. We formulate here a particular case of Theorem 4.1 from Section 4.

Let \( X \) be a compact subset of \( \mathbb{R}^d \). Let \( \mathcal{P}_n \) denote the set of all partitions of \( X \) into \( n \) disjoint Borel subsets. Let \( p_n \in \mathcal{P}_n, n = 1, \ldots \). Define \( L_n \) as a subspace of all functions that are piecewise constant on the partition \( p_n \). For a finite dimensional linear subspace \( L \subset L_2(\rho_X) \) and \( f \in L_2(\rho_X) \) we denote by \( d(f, L)_{L_2(\rho_X)} \) the \( L_2(\rho_X) \) distance between \( f \) and \( L \).
Theorem 1.7. Let \( \rho \) be such that \( |y| \leq M \) a.e. For a given sequence \( \{L_n\}_{n=1}^{\infty} \) and numbers \( m, r > 0, A \geq A_0(M,r) \) there exists an estimator \( f_z \) such that for any \( \rho \) we get

\[
d(f_{\rho}, L_n)_{L_2(\rho X)} \leq Dn^{-r}, \quad n = 1, 2, \ldots,
\]

we get

\[
\text{Prob}_{z \in \mathbb{Z}^m} \{\|f_{\rho} - f_z\|_{L_2(\rho X)}^2 \leq (1 + D^2)A(m \ln m/m)^{1/(1+2r)}\} \\
\geq 1 - \exp(-c(M)A(m \ln m)^{2r/(1+2r)}).
\]

Let us now discuss one more important issue. First, we remind the general scheme that we follow in constructing an estimator \( f_z \). We begin with a function class \( \mathcal{W} \). Then we look for an estimator that provides good estimation for the class \( \mathcal{W} \). In examples considered in Sections 2 and 3 we choose a hypothesis space \( \mathcal{H} \) where \( f_z \) comes from depending on the class \( \mathcal{W} \). It is a weak point of the above approach. In many cases we do not know exactly the class \( \mathcal{W} \). However, we may know a collection \( \mathcal{W} \) of classes where our unknown class \( \mathcal{W} \) belongs. Say, if we are thinking about \( \mathcal{W} \) in terms of Sobolev smoothness classes we may take as \( \mathcal{W} \) the collection of all Sobolev classes with smoothness from a certain range. We now discuss the universal method setting (see [DKPT]). In this setting a collection \( \mathcal{W} \) of classes is given and we need to find a procedure for constructing an estimator \( f_z \) in such a way that if \( f_{\rho} \in \mathcal{W} \) then \( \|f_{\rho} - f_z\|_{L_2(\rho X)} \) is close to the optimal error for the class \( \mathcal{W} \) with high probability with regard to \( \rho \times \cdots \times \rho \) (\( m \) times). In approximation theory this approach is known under the name of universal method (see [T1–T4]). We would like to build a universal estimator \( f_z \) for a given collection \( \mathcal{W} \) of classes. In Sections 4 and 5 we address this issue. We use different ideas in constructing universal estimators. In Section 4 we prove the following theorem.

Theorem 1.8. Let \( \rho \) be such that \( |y| \leq M \) a.e. For a given sequence \( \{L_n\}_{n=1}^{\infty} \) and numbers \( m, A \geq A_0(M) \) there exists an estimator \( f_z \) such that for some \( r \in (0,1/2] \) and some \( \rho \) we have

\[
d(f_{\rho}, L_n)_{L_2(\rho X)} \leq Dn^{-r},
\]

then

\[
\text{Prob}_{z \in \mathbb{Z}^m} \{\|f_{\rho} - f_z\|_{L_2(\rho X)} \leq C(D)A^{1/2}(m \ln m/m)^{1/(1+2r)}\} \\
\geq 1 - Cm^{-c(M)A}.
\]

We point out that the estimator \( f_z \) from Theorem 1.8 does not depend on both \( \rho_X \) and the specifics of \( \mathcal{W} \). This means that \( f_z \) is a universal estimator.

In Sections 2–4 we build estimators \( f_z \) as empirical optimums with hypothesis spaces \( \mathcal{H} \) suitable for a concrete problem under investigation. In constructing universal estimators in Section 4 we employ the following two ideas: 1. use the \( L_\infty \) balls of finite dimensional linear subspaces as hypothesis spaces; 2. minimise a penalized empirical risk. The above method uses the empirical risk function of the form

\[
\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2
\]
that is designed for measuring the approximation error \( \| f_z - f_\rho \| \) in the \( L_2(\rho_X) \) norm. In Section 5 we discuss a particular setting where we obtain the approximation error estimate in the \( L_\infty(\rho_X) \) norm. In this setting we assume that \( \rho_X \) is a normalized Lebesgue measure on a bounded domain \( \Omega \subset \mathbb{R}^d \). Next, we formulate our assumptions and build estimators in terms of a given sequence of kernels \( K_n \) of integral operators. A special case of \( K_n(x, u) = V_n(x-u) \) - the de la Vallée Poussin kernel, \( \Omega = [0, 2\pi] \), has been considered in [DKPT]. The technique used in Section 5 is a generalization of the corresponding technique from [DKPT].

We note that in [DKPT] the above setting with \( \rho_X \) the Lebesgue measure has been interpreted as a particular case of a general setting with estimating a function \( f_\mu \) instead of \( f_\rho \). In this setting we assume that \( \rho_X \) is an absolutely continuous measure with density \( \mu(x) \): \( d\rho_X = \mu dx \). We define \( f_\mu := f_\rho \mu \). Then we estimate \( f_\mu \) instead of \( f_\rho \). It is clear that in the case of \( \rho_X \) is the Lebesgue measure we have \( f_\mu = f_\rho \). One can find in [DKPT] a motivation for considering \( f_\mu \).

In Section 5 we build an estimator for \( f_\rho \) by the formula

\[
f_z := \frac{1}{m} \sum_{i=1}^{m} y_i K_n(x, x_i)
\]

which is simpler than an empirical optimum. In constructing a universal estimator instead of penalization we use the size of the corresponding dyadic blocks

\[
f_{s,z} := \frac{1}{m} \sum_{i=1}^{m} y_i (K_{2^s}(x, x_i) - K_{2^{s-1}}(x, x_i)).
\]

2. The case \( r > 1/2 \)

In the case of restrictions imposed in the uniform norm \( \mathcal{C} \) the following theorem has been proved in [KT] (see Theorem 1.3 from Introduction). We reformulate it here for convenience.

**Theorem 2.1 [KT].** Assume that \( \rho, W \) satisfy (1.5) and \( W \) is such that

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

Then for \( m\eta^2 \geq 1 \) we have

\[
\text{Prob}_{z \in \mathbb{Z}^m \{ \sup_{f \in W} |L_z(f)| \geq \eta \} \leq C(M, \epsilon(W)) \exp(-c(M)m\eta^2)}.
\]

First of all we will show that Theorem 2.1 cannot be extended onto the case \( L_2(\rho_X) \) in its form. The following example shows that if we consider entropy of \( W \) in \( L_2(\rho_X) \) rather than in \( \mathcal{C}[0,1] \) then even a fast decay of \( \epsilon_n(W, L_2(\rho_X)) \) (say, \( \epsilon_n(W, L_2(\rho_X)) = o(n^{-r}) \) for every \( r > 0 \)) does not guarantee nontrivial estimates for \( \sup_{f \in W} |L_z(f)| \). We assume that \( Y = [-1, 1] \), and thus, the functions \( f \in W \) and \( f_\rho \) are uniformly bounded.
Proposition 2.1. Let $N$ be a non-increasing mapping $(0, +\infty) \rightarrow [1, +\infty)$ such that
\[ \lim_{u \to 0^+} \frac{\log N(u)}{\log(1/u)} = +\infty. \]
Then there exist a set $W \subset U(L_\infty[0,1])$ and a $\rho$ such that
\[ N(W, \epsilon, L_2(\rho_X)) \leq N(\epsilon) \]
and for every $m$
\[ \text{Prob}_{z \in Z^m}{\{\sup_{f \in W} |L_z(f)| \leq 1/2\}} = 0. \]

Proof. Let us take an increasing sequence $\{K_m\}$ of positive integers so that
\[ K_m > 2m^3, \quad N(K_m^{-1}/3) \geq K_m^{m+1} \quad (m \in \mathbb{N}). \]
The existence of $K_m$ satisfying (2.4) follows from our assumption (2.2). For every $m$, every $l = (l_1, \ldots, l_m)$, $1 \leq l_1 < \cdots < l_m \leq K_m$, and every $x \in [0,1)$ we define
\[ f_{m,l}(x) = \begin{cases} 1, & \text{if } [K_mx] + 1 \in \{l_1, \ldots, l_m\}, \\ 0, & \text{otherwise}. \end{cases} \]
Let $W_m = \{f_{m,l}\}$, $f_0 \equiv 0$, $W = \{f_0\} \cup \bigcup_m W_m$. We denote $\epsilon_m = K_m^{-1/3}$. By (2.4), for any $f \in W_m$ we have
\[ \|f\|_{L_2[0,1]} \leq (m/K_m)^{1/2} \leq \epsilon_m \]
and also
\[ \|f\|_{L_2[0,1]}^2 < 1/2. \]
Let us check (2.3). If $\epsilon \geq \epsilon_1$, then $\{f_0\}$ forms a $\epsilon_1$-net in the $L_2[0,1]$ norm, and (2.3) holds since $N(\epsilon) \geq 1$. If $\epsilon < \epsilon_1$, then we can find $m$ so that $\epsilon_{m+1} \leq \epsilon < \epsilon_m$ and using (2.5) take the following $\epsilon$-net for $W$:
\[ A = \{f_0\} \cup \bigcup_{j \leq m} W_j. \]
We have
\[ \#A \leq 1 + \sum_{j=1}^m \#W_j \leq 1 + \sum_{j=1}^m K_j^j \leq (m + 1)K_m^m < K_m^{m+1}, \]
and, by (2.4),
\[ \#A < N(\epsilon_m) \leq N(\epsilon). \]
So, (2.3) holds.
We now take $\rho$ so that $\rho_X$ is the Lebesgue measure on $[0,1)$ and $y$ is surely 0 for any $x$. Clearly, $f_\rho \equiv 0$. On the one hand, by (2.6), we have for any $f \in W$
\[ \mathcal{E}(f) < 1/2. \]
On the other hand, for any $z$ there is $f \in W_m$ so that $f(x_i) = 1 (i = 1, \ldots, m)$. Therefore, $\mathcal{E}_z(f) = 1$, $L_z(f) < -1/2$, and Proposition 2.1 is proven.

We will prove an analog of Theorem 2.1 in the case of $L_2(\rho_X)$ norm with the set $W$ replaced by a $\delta$-net $\mathcal{N}_\delta(W)$ of $W$ in the $L_2(\rho_X)$ norm. We begin with an auxiliary lemma.
Lemma 2.1. If \(|f_j(x) - y| \leq M\) a.e. for \(j = 1, 2\) and \(\|f_1 - f_2\|_{L_2(\rho_X)} \leq \delta\), then for \(\delta^2 \geq \eta\)
\[
\text{Prob}_{z \in Z^m}\{|L_z(f_1) - L_z(f_2)| \leq \eta\} \geq 1 - 2\exp\left(-\frac{m\eta^2}{9M^2\delta^2}\right).
\]
and for \(\delta^2 < \eta\)
\[
\text{Prob}_{z \in Z^m}\{|L_z(f_1) - L_z(f_2)| \leq \eta\} \geq 1 - 2\exp\left(-\frac{m\eta}{9M^2}\right).
\]

Proof. Consider the random variable \(\xi = (f_1(x) - y)^2 - (f_2(x) - y)^2\). We use
\(|\xi| \leq M^2\), \(\sigma(\xi) \leq 2M\delta\).

Applying the Bernstein inequality (1.4) to \(\xi\) we get
\[
\text{Prob}_{z \in Z^m}\{|L_z(f_1) - L_z(f_2)| \geq \eta\} = \text{Prob}_{z \in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)\right| \geq \eta\right\} \leq 2\exp\left(-\frac{m\eta^2}{2(4M^2\delta^2 + M^2\eta/3)}\right),
\]
and Lemma 2.1 follows.

Theorem 2.2. Assume that \(\rho, W\) satisfy (1.5) and \(W\) is such that
\[
\sum_{n=1}^{\infty} n^{-1/2}\epsilon_n(W, L_2(\rho_X)) < \infty.
\]

Let \(m\eta^2 \geq 1\). Then for any \(\delta\) satisfying \(\delta^2 \geq \eta\) we have for a minimal \(\delta\)-net \(\mathcal{N}_\delta(W)\) of \(W\) in the \(L_2(\rho_X)\) norm
\[
\text{Prob}_{z \in Z^m}\{\sup_{f \in \mathcal{N}_\delta(W)} |L_z(f)| \geq \eta\} \leq C(M, \epsilon(W)) \exp(-c(M)m\eta^2).
\]

Proof. It is clear that (2.7) implies that
\[
\sum_{j=0}^{\infty} 2^{j/2}\epsilon_{2^j}(W, L_2(\rho_X)) < \infty.
\]

Denote \(\delta_j := \epsilon_{2^j}(W, L_2(\rho_X))\), \(j = 0, 1, \ldots\), and consider minimal \(\delta_j\)-nets \(\mathcal{N}_\delta := \mathcal{N}_\delta(W) \subset W\) of \(W\). We will use the notation \(\mathcal{N}_j := |\mathcal{N}_j|\). Let \(J\) be the minimal \(j\) satisfying \(\delta_j \leq \delta\). We modify \(\delta_j\) by setting \(\delta_J = \delta\). Then \(\mathcal{N}_J = \mathcal{N}_{\delta}(W)\). For \(j = 1, \ldots, J\) we define a mapping
consider the case $A_j$ that associates with a function $f \in W$ a function $A_j(f) \in \mathcal{N}_j$ closest to $f$ in the $L_2(\rho_X)$ norm. Then, clearly,

$$\|f - A_j(f)\|_{L_2(\rho_X)} \leq \delta_j.$$ 

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

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

We introduce an auxiliary sequence

$$\eta_j := 3M\eta^{2(j+1)/2}2^{j-1}, \quad j = 1, 2, \ldots,$$ 

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

$$\sum_{j \geq I} M2^{(j+1)/2}2^{j-1} \leq 1/6 \quad \text{or} \quad \sum_{j \geq I} \eta_j \leq \eta/2.$$ 

We now proceed to the estimate of $\text{Prob}_{z \in \mathbb{Z}^m}\{\sup_{f \in \mathcal{N}_i(W)} |L_z(f)| \geq \eta\}$ with $m, \eta$ satisfying $m\eta^2 \geq 1$. If $J \leq I$ then the statement of Theorem 2.2 follows from Theorem 1.2. We consider the case $J > I$. Assume $|L_z(f_J)| \geq \eta$. Then rewriting

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

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

$$|L_z(f_J) - L_z(f_{J-1})| \geq \eta_j \quad \text{for some} \quad j \in (I, J) \quad \text{or} \quad |L_z(f_{I})| \geq \eta/2.$$ 

Therefore

$$\text{Prob}_{z \in \mathbb{Z}^m}\{\sup_{f \in \mathcal{N}_i(W)} |L_z(f)| \geq \eta\} \leq \text{Prob}_{z \in \mathbb{Z}^m}\{\sup_{f \in \mathcal{N}_i} |L_z(f)| \geq \eta/2\}$$

$$+ \sum_{j \in (I, J)} \sum_{f \in \mathcal{N}_j} \text{Prob}_{z \in \mathbb{Z}^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\}$$

$$\leq \text{Prob}_{z \in \mathbb{Z}^m}\{\sup_{f \in \mathcal{N}_i} |L_z(f)| \geq \eta/2\}$$

$$+ \sum_{j \in (I, J)} N_j \sup_{f \in \mathcal{W}} \text{Prob}_{z \in \mathbb{Z}^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\}.$$ 

By our choice of $\delta_j = \epsilon_{2j}(W, L_2(\rho_X))$ we get $N_j \leq 2^{2j} < e^{2j}$. Let $\eta, \delta$ be such that $m\eta^2 \geq 1$ and $\eta \leq \delta^2$. It is clear that $\delta_j^2 \geq \eta_j$, $j = 1, \ldots, J$. Applying Lemma 2.1 we obtain

$$\sup_{f \in \mathcal{W}} \text{Prob}_{z \in \mathbb{Z}^m}\{|L_z(f) - L_z(A_{j-1}(f))| \geq \eta_j\} \leq 2 \exp\left(-\frac{m\eta^2}{9M^2\delta_j^2}\right), \quad j \leq J.$$
From the definition (2.9) of \( \eta_j \) we get
\[
\frac{m \eta_j^2}{9M^2 \delta_j^2} = m \eta^2 2^{j+1}
\]
and
\[
N_j \exp \left( -\frac{m \eta_j^2}{9M^2 \delta_j^2} \right) \leq \exp(-m \eta^2 2^j).
\]
Therefore
\[
\sum_{j \in (I,J]} N_j \exp \left( -\frac{m \eta_j^2}{9M^2 \delta_j^2} \right) \leq 2 \exp(-m \eta^2 2^I).
\]
By Theorem 1.2
\[
\text{Prob}_{z \in Z^m} \left\{ \sup_{f \in \mathcal{N}_I} |L_z(f)| \geq \eta/2 \right\} \leq 2N_I \exp \left( -\frac{m \eta^2}{C(M)} \right).
\]
Combining (2.12) and (2.13) we obtain
\[
\text{Prob}_{z \in Z^m} \left\{ \sup_{f \in \mathcal{N}_I(W)} |L_z(f)| \geq \eta \right\} \leq C(M, \epsilon(W)) \exp(-c(M)m \eta^2).
\]
This completes the proof of Theorem 2.2.

We get the following error estimates for \( \mathcal{E}(f_z) - \mathcal{E}(f_W) \) from Theorem 2.2.

**Theorem 2.3.** Assume that \( \rho, W \) satisfy (1.5), (2.7), and also \( f_\rho \in W \). Let \( m \eta^2 \geq 1 \). Then there exists an estimator \( f_z \) such that
\[
\text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3\eta \} \geq 1 - C(M, \epsilon(W)) \exp(-c(M)m \eta^2)
\]
with \( C(M, \epsilon(W)), c(M) \) from Theorem 2.2.

**Proof.** Let us take \( \delta = \eta^{1/2} \) and \( \mathcal{H} := \mathcal{N}_\delta(W) \) a minimal \( \delta \)-net for \( W \) in the \( L_2(\rho_X) \) norm, \( f_z = f_z, \mathcal{H} \). Then we have \( (f_W = f_\rho) \)

\[
\mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_W) = \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_W) + \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_z, \mathcal{H}) + \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_z, \mathcal{H}) + \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\mathcal{H})
\]
\[
+ \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_\mathcal{H}) \leq \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_W) + \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_z, \mathcal{H}) + \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\mathcal{H}) + \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\mathcal{H}).
\]

Therefore,
\[
\mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_W) \leq \eta + \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_z, \mathcal{H}) + \mathcal{E}(f_\mathcal{H}) - \mathcal{E}(f_\mathcal{H}),
\]
and to complete the proof it remains to use Theorem 2.2.

Let us now prove an estimate for \( \mathcal{E}(f_z) - \mathcal{E}(f_W) \) without an assumption \( f_\rho \in W \).
Theorem 2.4. Assume that \( \rho, W \) satisfy (1.5), (1.2) with \( r > 1/2 \). Let \( m\eta^{1+\max(1/r,1)} \geq A_0(M,D,r) \geq 1 \). Then there exists an estimator \( f_z \in W \) such that

\[
\text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_W) \leq 5\eta \} \geq 1 - C_1(M,D,r) \exp(-c_1(M)m\eta^2).
\]

Proof. It suffices to prove the theorem for \( r \in (1/2,1] \). Let us take \( \delta_0 := \eta^{1/2} \) and \( \mathcal{H}_0 := \mathcal{N}_{\delta_0}(W) \) to be a minimal \( \delta_0 \)-net for \( W \). Let \( \delta := \eta/(2M) \) and \( \mathcal{H} := \mathcal{N}_{\delta}(W) \) to be a minimal \( \delta \)-net for \( W \). Denote \( f_z := f_{z,H} \). For any \( f \in \mathcal{H} \) there is \( A(f) \in \mathcal{H}_0 \) such that \( \|f - A(f)\|_{L^2(\rho_X)} \leq \delta_0 \). By Lemma 2.1,

\[
\text{Prob}_{z \in Z^m} \{ |L_z(f) - L_z(A(f))| \leq \eta \} \geq 1 - 2 \exp \left(-\frac{m\eta}{9M^2}\right).
\]

Using the above inequality and Theorem 2.2 \( (m\eta^2 \geq 1) \) we get

\[
(2.16) \quad \text{Prob}_{z \in Z^m} \{ \sup_{f \in \mathcal{H}} |L_z(f)| \geq 2\eta \} \leq \text{Prob}_{z \in Z^m} \{ \sup_{f \in \mathcal{H}} |L_z(f) - L_z(A(f))| \geq \eta \}
\]

\[
+ \text{Prob}_{z \in Z^m} \{ \sup_{f \in \mathcal{H}_0} |L_z(f)| \geq \eta \} \leq 2\#\mathcal{H} \exp \left(-\frac{m\eta}{9M^2}\right) + C(M,D,r) \exp(-c(M)m\eta^2)
\]

\[
\leq 4 \exp \left((\eta^{-1/r}(2MD)^{1/r})\right) \exp \left(-\frac{m\eta}{9M^2}\right) + C(M,D,r) \exp(-c(M)m\eta^2).
\]

Let us specify \( A_0(M,D,r) := \max(18M^2(2MD)^{1/r},1) \), \( r \in (1/2,1] \). Then

\[
(2.17) \quad m\eta^{1+1/r} \geq 18M^2(2MD)^{1/r}
\]

and (2.16) imply

\[
\text{Prob}_{z \in Z^m} \{ \sup_{f \in \mathcal{H}} |L_z(f)| \geq 2\eta \} \leq 4 \exp \left(-\frac{m\eta}{18M^2}\right) + C(M,D,r) \exp(-c(M)m\eta^2).
\]

Further, we can assume that \( \eta < M^2 \) (otherwise, the statement of Theorem 2.4 is trivial). Therefore, we deduce from the last estimate that

\[
\text{Prob}_{z \in Z^m} \{ \sup_{f \in \mathcal{H}} |L_z(f)| \geq 2\eta \} \leq C_1(M,D,r) \exp(-c_1(M)m\eta^2).
\]

We now observe that, by the choice of \( \delta \),

\[
(2.18) \quad \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_W) = \|f_{\mathcal{H}} - f_{\rho}\|^2_{L^2(\rho_X)} - \|f_W - f_{\rho}\|^2_{L^2(\rho_X)}
\]

\[
= (\|f_{\mathcal{H}} - f_{\rho}\|^2_{L^2(\rho_X)} - \|f_W - f_{\rho}\|^2_{L^2(\rho_X)}) \leq \eta.
\]

Using (2.14) we see that (2.15) holds. Hence, if \( \sup_{f \in \mathcal{H}} |L_z(f)| \leq 2\eta \), then \( \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_W) \leq 5\eta \). This completes the proof of Theorem 2.4.

Theorem 2.5. Let \( f_{\rho} \in W \) and let \( \rho, W \) satisfy (1.5) and (1.2) with \( r > 1/2 \). Then there exists an estimator \( f_z \) such that for \( A \geq 2 \)

\[
(2.19) \quad \text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_{\rho}) \leq 3A^{1/2}(\ln m/m)^{1/2} \} \geq 1 - C(M,D,r)m^{-c(M)A}.
\]

Also

\[
(2.20) \quad \text{Prob}_{z \in Z^m} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_{\rho})| \leq 4A^{1/2}(\ln m/m)^{1/2} \} \geq 1 - C(M,D,r)m^{-c(M)A}.
\]

Proof. First, we use Theorem 2.3 with \( \eta = A^{1/2}(\ln m/m)^{1/2} \) and get (2.19) with \( f_z = f_{z,\mathcal{N}_{\eta,1/2}(W)} \). Second, we use Theorem 2.2 with the above \( \eta \) and \( \delta = \eta^{1/2} \) and obtain (2.20).
3. The case \( r \in (0, 1/2) \)

The following results have been obtained in [KT] in the case when we impose restrictions in the uniform norm \( C \).

**Theorem 3.1 [KT].** Assume that \( \rho, W \) satisfy (1.5) and \( W \) is such that

\[
\sum_{n=1}^{\infty} n^{-1/2} \epsilon_n = \infty, \quad \epsilon_n := \epsilon_n(W, C).
\]

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

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

Then for \( m, \eta \) satisfying \( m \eta^2 \geq 480M^2 \) we have

\[
\Pr_{z \in Z^m} \{ \sup_{f \in W} |L_z(f)| \geq \eta \} \leq C(M, \epsilon(W)) \exp(-c(M)m(\eta/S_J)^2).
\]

**Corollary 3.1 [KT].** Assume \( \rho, W \) satisfy (1.5) and \( \epsilon_n(W, C) \leq Dn^{-1/2} \). Then for \( m, \eta \) satisfying \( m \eta^2/(1 + (\log(M/\eta))^2) \geq C_1(M, D) \) we have

\[
\Pr_{z \in Z^m} \{ \sup_{f \in W} |L_z(f)| \geq \eta \} \leq C(M, D) \exp(-c(M, D)m \eta^2/(1 + (\log(M/\eta))^2)).
\]

**Corollary 3.2 [KT].** Assume \( \rho, W \) satisfy (1.5) and \( \epsilon_n(W, C) \leq Dn^{-r}, \ r \in (0, 1/2) \). Then for \( m, \eta, \delta \geq \eta/(8M) \) satisfying \( m \eta^2 \delta^{1/r-2} \geq C_1(M, D, r) \) we have

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

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

We prove here the following analogs of these results with restrictions imposed in the \( L_2(\rho_X) \) norm.

**Theorem 3.2.** Assume that \( \rho, W \) satisfy (1.5) and

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

Let \( \eta, \delta \) be such that \( \delta^2 \geq \eta \). Define \( J := J(\delta) \) as the minimal \( j \) satisfying \( \epsilon_{2j} \leq \delta \) and

\[
S_J := \sum_{j=1}^{J} 2^{(j+1)/2} \epsilon_{2j-1}, \quad J \geq 1; \quad S_0 := 1.
\]
Then for $m$, $\eta$ satisfying $m(\eta/S_J)^2 \geq 36M^2$ we have
$$\text{Prob}_{z \in \mathbb{Z}^m} \left\{ \sup_{f \in \mathcal{N}_\delta(W)} \left| L_z(f) \right| \geq \eta \right\} \leq C(M, \epsilon(W)) \exp(-c(M)m(\eta/S_J)^2),$$
where $\mathcal{N}_\delta(W)$ is a minimal $\delta$-net of $W$ in the $L_2(\rho_X)$.

**Proof.** In the case $J = 0$ the statement of Theorem 3.2 follows from Theorem 1.1. In the case $J \geq 1$ the proof differs from the proof of Theorem 2.2 only in the choice of an auxiliary sequence $\{\eta_j\}$. Thus we keep notations from the proof of Theorem 2.2. Now, instead of (2.9) we define $\{\eta_j\}$ as follows
$$\eta_j := \frac{\eta 2^{(j+1)/2} \epsilon_{2j-1}}{S_J}.$$ Proceeding as in the proof of Theorem 2.2 with $I = 1$ we need to check that
$$2^j - \frac{mn_j^2}{9M^2 \delta_{j-1}^2} \leq -2^j m(\eta/S_J)^2 \frac{\eta}{36M^2}.$$ Indeed, using the assumption $m(\eta/S_J)^2 \geq 36M^2$ we obtain
$$\frac{mn_j^2}{9M^2 \delta_{j-1}^2} - 2^j = \frac{m(\eta/S_J)^2}{36M^2} 2^{j+1} - 2^j \geq \frac{m(\eta/S_J)^2}{36M^2} 2^j.$$ We complete the proof in the same way as in Theorem 2.2.

**Corollary 3.3.** Assume $\rho$, $W$ satisfy (1.5) and $\epsilon_n(W, L_2(\rho_X)) \leq Dn^{-1/2}$. Then for $m$, $\eta$ satisfying $mn_j^2/(1 + (\log(M/\eta))^2) \geq C_1(M, D)$ we have for $\delta^2 \geq \eta$
$$\text{Prob}_{z \in \mathbb{Z}^m} \left\{ \sup_{f \in \mathcal{N}_\delta(W)} \left| L_z(f) \right| \geq \eta \right\} \leq C(M, D) \exp(-c(M, D)m\eta^2/(1 + (\log(M/\eta))^2)).$$

**Corollary 3.4.** Assume $\rho$, $W$ satisfy (1.5) and $\epsilon_n(W, L_2(\rho_X)) \leq Dn^{-r}$, $r \in (0, 1/2)$. Then for $m$, $\eta$, $\delta^2 \geq \eta$ satisfying $mn_j^2 \delta^{1/r-2} \geq C_1(M, D, r)$ we have
$$\text{Prob}_{z \in \mathbb{Z}^m} \left\{ \sup_{f \in \mathcal{N}_\delta(W)} \left| L_z(f) \right| \geq \eta \right\} \leq C(M, D, r) \exp(-c(M, D, r)m\eta^2 \delta^{1/r-2}).$$

The proofs of both corollaries are the same. We present here only the proof of Corollary 3.4.

**Proof of Corollary 3.4.** We use Theorem 3.2. Similarly to the proof of Theorem 3.2 it is sufficient to consider the case $J \geq 1$. We estimate the $S_J$ from Theorem 3.2:
$$S_J = \sum_{j=1}^{J} 2^{(j+1)/2} \epsilon_{2j-1} \leq 2^{1/2+r} D \sum_{j=1}^{J} 2^{j(1/2-r)} \leq C_1(r) D 2^{J(1/2-r)}.$$ Next,
$$D 2^{-r(J-1)} \geq \epsilon_{2j-1} \geq \delta \implies 2^J \leq 2(D/\delta)^{1/r}.$$ Thus
$$S_J \leq C_1(D, r)(1/\delta)^{\frac{1}{r}}.$$ It remains to apply Theorem 3.2.
Theorem 3.3. Let \(f_\rho \in W\) and let \(\rho, W\) satisfy (1.5) and (1.2). Then there exists an estimator \(f_z\) such that

\[
\text{Prob}_{z \in Z} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3\eta \} \geq 1 - C(M, D) \exp(-c(M, D) m \eta^2 / (1 + (\log(M/\eta))^2)),
\]

(3.2) \(\text{Prob}_{z \in Z} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_\rho)| \leq 4\eta \} \geq 1 - C(M, D) \exp(-c(M, D) m \eta^2 / (1 + (\log(M/\eta))^2)),\)

provided \(r = 1/2\), \(m \eta^2 / (1 + (\log(M/\eta))^2) \geq C_1(M, D),\)

(3.3) \(\text{Prob}_{z \in Z} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3\eta \} \geq 1 - C(M, D, r) \exp(-c(M, D, r) m \eta^{1+1/(2r)}),\)

(3.4) \(\text{Prob}_{z \in Z} \{ |\mathcal{E}(f_z) - \mathcal{E}(f_\rho)| \leq 4\eta \} \geq 1 - C(M, D, r) \exp(-c(M, D, r) m \eta^{1+1/(2r)}),\)

provided \(r \in (0, 1/2), m \eta^{1+1/(2r)} \geq C_1(M, D, r)\) with constants \(C(M, D), c(M, D), C_1(M, D), C(M, D, r), c(M, D, r), C_1(M, D, r)\) from Corollaries 3.3 and 3.4.

Proof. We combine the proof of Theorem 2.3 with Corollaries 3.3 and 3.4. In the case \(r = 1/2\) we take \(\eta\) such that \(m \eta^2 / (1 + (\log(M/\eta))^2) \geq C_1(M, D)\) and set \(\delta = \eta^{1/2}\). Denote \(\mathcal{H} := \mathcal{N}_\delta(W)\). Then similarly to (2.14), (2.15) we obtain

\[
\mathcal{E}(f_{z, \mathcal{H}}) - \mathcal{E}(f_\rho) \leq \delta^2 + \mathcal{E}(f_{z, \mathcal{H}}) - \mathcal{E}(f_{z, \mathcal{H}}) + \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}).\]

Using Corollary 3.3 we continue

\[
\leq 3\eta
\]

with probability at least \(1 - C(M, D) \exp(-c(M, D) m \eta^2 / (1 + (\log(M/\eta))^2))\). This proves (3.1). Applying Corollary 3.3 one more time we obtain (3.2).

We proceed to the case \(r \in (0, 1/2)\). We now take \(\eta\) such that \(m \eta^{1+1/(2r)} \geq C_1(M, D, r)\) and set \(\delta = \eta^{1/2}\). Denote as above \(\mathcal{H} := \mathcal{N}_\delta(W)\). We now use (3.5) and apply Corollary 3.4. We get

\[
\mathcal{E}(f_{z, \mathcal{H}}) - \mathcal{E}(f_\rho) \leq \delta^2 + 2\eta \leq 3\eta
\]

with probability at least

\[
1 - C(M, D, r) \exp(-c(M, D, r) m \eta^{1+1/(2r)}).
\]

This proves (3.3). Applying Corollary 3.4 again we get (3.4). The proof of Theorem 3.3 is now complete.

We give a direct corollary of Theorem 3.3.
Corollary 3.5. Let \( f_\rho \in W \) and let \( \rho, W \) satisfy (1.5) and (1.2). Then there exists an estimator \( f_z \) such that for \( A \geq A_0(M,D,r) \geq 1 \)

\[
\operatorname{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3A((\ln m)^3/m)^{1/2} \} \geq 1 - C(M,D)m^{-c(M,D)A^2},
\]

\[
\operatorname{Prob}_{z \in Z^m} \{ |\mathcal{E}_z(f_z) - \mathcal{E}(f_\rho)| \leq 4A((\ln m)^3/m)^{1/2} \} \geq 1 - C(M,D)m^{-c(M,D)A^2},
\]

provided \( r = 1/2 \),

\[
\operatorname{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq 3A(\ln m/m)^{\frac{2r}{1+2r}} \} \geq 1 - C(M,D,r)m^{-c(M,D,r)A^{1+\frac{1}{2r}}},
\]

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

for \( m \geq C(A,M) \) provided \( r \in (0,1/2) \) with constants \( C(M,D), c(M,D), C(M,D,r), c(M,D,r) \) from Corollaries 3.3 and 3.4.

We now prove an analog of Theorem 2.4.

Theorem 3.4. Assume that \( \rho, W \) satisfy (1.5), (1.2) with \( r \in (0,1/2] \). Let \( mn^{1+1/r} \geq A_0(M,D,r) \geq 1 \). Then there exists an estimator \( f_z \in W \) such that

\[
\operatorname{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_W) \leq 5\eta \} \geq 1 - C(M,D) \exp(-c(M,D)mn^2/(1 + (\log(M/\eta))^2))
\]

provided \( r = 1/2 \),

\[
\operatorname{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z) - \mathcal{E}(f_W) \leq 5\eta \} \geq 1 - C(M,D,r) \exp(-c(M,D,r)mn^{1+1/(2r)})
\]

provided \( r \in (0,1/2) \).

Proof. The proof in both cases \( r = 1/2 \) and \( r \in (0,1/2) \) is similar to the proof of Theorem 2.4. We will sketch the proof only in the case \( r \in (0,1/2) \), \( \eta \leq 1 \). We use the notations from the proof of Theorem 2.4. We choose \( A_0(M,D,r) \geq C_1(M,D,r) \) - the constant from Corollary 3.4. Then we can use Corollary 3.4 with \( \delta = \eta^{1/2} \) because

\[
m\eta^2\delta^{1/r-2} = m\eta^{1+1/(2r)} \geq m\eta^{1+1/r} \geq A_0(M,D,r) \geq C_1(M,D,r).
\]

We obtain the following analog of (2.16)

\[
\operatorname{Prob}_{z \in Z^m} \left\{ \sup_{f \in \mathcal{H}} |L_z(f)| \geq 2\eta \right\}
\]

\[
\leq 4 \exp \left( (\eta^{-1/r})(2MD)^{1/r} \right) \exp \left( -\frac{mn^2}{9M^2} \right) + C(M,D,r) \exp(-c(M,D,r)mn^{1+1/(2r)}).
\]

We complete the proof in the same way as in the proof of Theorem 2.4.
4. Some specifications

Assume that \( n \)-dimensional linear subspaces \( L_n \) have the following property: for any probability measure \( w \) on \( X \) one has

\[
\|P^w_{L_n}\|_{L_\infty(w) \to L_\infty(w)} \leq K, \quad n = 1, 2, \ldots
\]

where \( P^w_L \) is the operator of \( L_2(w) \) projection onto \( L \). First of all we note that

\[
d(f_\rho, L_n)_{L_2(\rho_X)} = \|f_\rho - P^\rho_X L_n(f_\rho)\|_{L_2(\rho_X)}.
\]

In this section we will assume that \( |y| \leq M \) a.e. Then by (4.1) we get

\[
\|P^\rho_X L_n(f_\rho)\|_{L_\infty(\rho_X)} \leq MK.
\]

Denote \( V_n := MKU(L_\infty(\rho_X)) \cap L_n \).

**Theorem 4.1.** Let \( \rho \) be such that \( |y| \leq M \) a.e. Assume that a sequence \( \{L_n\}_{n=1}^\infty \) satisfies (4.1). For given \( m, r > 0, A \geq A_0(M, K, r) \) there exists an estimator \( f_z \) such that for any \( \rho \) satisfying

\[
d(f_\rho, L_n)_{L_2(\rho_X)} \leq Dn^{-r}, \quad n = 1, 2, \ldots,
\]

we get

\[
\text{Prob}_{z \in Z^m}\{\|f_\rho - f_z\|^2_{L_2(\rho_X)} \leq (1 + D^2)A(\ln m/m)^{2r} \} \geq 1 - \exp(-c(M)A(\ln m)^{2r})^{\frac{1}{1+2r}}.
\]

**Proof.** We set \( \epsilon = A(\ln m/m)^{\frac{2r}{1+2r}}, n = [\epsilon^{-1/(2r)}] + 1 \) and \( f_z := f_{z,V_n} \). We now estimate \( \mathcal{E}(f_{z,V_n}) - \mathcal{E}(f_\rho) \). Let \( f^* := P^\rho_X L_n(f_\rho) \). Then by (4.1) \( f^* \in V_n \) and

\[
\|f_\rho - f^*\|_{L_2(\rho_X)} \leq Dn^{-r} \leq DA^{1/2}(\ln m/m)^{\frac{r}{1+2r}}.
\]

Therefore,

\[
\mathcal{E}(f^*) - \mathcal{E}(f_\rho) = \int_X (f^*(x) - f_\rho(x))^2 d\rho_X \leq D^2A(\ln m/m)^{2r}.
\]

We have

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

Denote for a compact subset \( \mathcal{H} \) of \( L_2(\rho_X) \)

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

It is clear that \( f^* = f_{V_n} \). We will use the following theorem from [CS].
Theorem 4.2 [CS]. Suppose that either $\mathcal{H}$ is a compact and convex subset of $L_\infty(\rho_X)$ or $\mathcal{H}$ is a compact subset of $L_\infty(\rho_X)$ and $f_\rho \in \mathcal{H}$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \leq M$ a.e. Then, for all $\epsilon > 0$

$$\text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f_\mathcal{H}) \leq \epsilon \} \geq 1 - N(\mathcal{H}, \epsilon/(24M), L_\infty(\rho_X))2\exp\left(-\frac{m\epsilon}{288M^2}\right).$$

It is well known that [P,p.63]

$$N(V_n, \epsilon, L_\infty(\rho_X)) \leq (1 + 2MK/\epsilon)^n.$$ Using this estimate and taking into account the choice of $\epsilon = A_0(M, K, r)$, we get from Theorem 4.2 for $A > A_0(M, K, r)$

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

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

Using (4.2) we obtain from here

$$\text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z, V_n) - \mathcal{E}(f_\rho) \leq (1 + D^2)A(\ln m/m)^{2r/(1+2r)} \}$$

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

This completes the proof of Theorem 4.1.

We note that the estimator $f_z = f_z, V_n$ from Theorem 4.1 does not depend on $\rho_X$ and depends on the class $W$ ($n$ is chosen using $r$). We will formulate one result on construction of universal estimators $f_z$ in a spirit of Theorem 2.6 from [DKPT]. For a given sequence $L = \{L_n\}_{n=1}^\infty$ satisfying (4.1) and for a given $m$ we define an estimator $f_z$ by the formula

$$f_z := f_z, V_k$$

with

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

Theorem 4.3. Assume that $L$ satisfies (4.1) and $\rho$ is such that $|y| \leq M$ a.e. Then if for some $r \in (0, 1/2]$

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

then we have

$$\text{Prob}_{z \in Z^m} \{ \|f_\rho - f_z\|_{L_2(\rho_X)} \leq C(D)A^{1/2}(\ln m/m)^{1/(1+2r)} \}$$

$$\geq 1 - Cm^{-c(M)A}, \quad A \geq A_0(M, K).$$

The proof of this theorem is similar to the proof of Theorem 2.6 from [DKPT].

Proof. We will use the following result from [CS] (it is a direct corollary to Proposition 7 from [CS]).
Lemma 4.1. Let $\mathcal{H}$ be a compact and convex subset of $L_\infty(\rho_X)$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \leq M$ a.e. Then for all $\epsilon > 0$ with probability at least

$$1 - N(\mathcal{H}, \frac{\epsilon}{24M}, L_\infty(\rho_X)) \exp\left(-\frac{me}{288M^2}\right)$$

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

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

By Bernstein’s inequality (1.4) we have

$$\text{Prob}_{z \in \mathbb{Z}^m}\{ \max_{1 \leq n \leq m} (\mathcal{E}(f_{V_n}) - \mathcal{E}_z(f_{V_n})) \leq A(\ln m/m)^{1/2} \} \geq 1 - 2m^{-c(M)A}.$$ 

Applying Lemma 4.1 with $\mathcal{H} = V_n$, $\epsilon = An \ln m/m$, $f = f_{z,V_n}$ and using that $\mathcal{E}(f_{V_n}) \geq \mathcal{E}(f_\rho)$ we get for $n \in [1, m]$, $A \geq A_0(M, K)$

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

with probability at least $1 - Cm^{-c(M)A}$. Therefore, for these $z$

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

We estimate $\min_{n \in [1, m]} 2\mathcal{E}_z(f_{z,V_n}) + An \ln m/m$ by the value at $n = n(r) := [(m/ \ln m)^{\frac{1}{1+2\gamma}}] + 1$. We have

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

Similarly to (4.7) we get

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

with probability $\geq 1 - 2m^{-c(M)A}$. Next,

$$\mathcal{E}(f_{V_{n(r)}}) - \mathcal{E}(f_\rho) = \| f_{V_{n(r)}} - f_\rho \|^2_{L_2(\rho_X)}$$

$$= d(f_\rho, L_{n(r)})^2 \leq D^2n(r)^{-2r} \leq D^2(\ln m/m)^{\frac{2r}{1+2\gamma}}.$$ 

Combining the relations (4.8)–(4.11) we obtain

$$\mathcal{E}(f_z) - \mathcal{E}(f_\rho) \leq C(D)A(\ln m/m)^{\frac{2r}{1+2\gamma}}.$$
with probability at least \(1 - Cm^{-c(M)A}\) provided \(A \geq A_0(M, K)\).

We now proceed to the case where we impose weaker than (4.5) restrictions on the class \(W\). These new restrictions are in a style of nonlinear Kolmogorov widths used in [DKPT] (see [T5]). Denote for a given \(a > 0\) \(N_n := [n^{an}]\). Let \(\mathcal{L}_n(a)\) be a collection of \(N_n\) dimensional subspaces \(L_n^1, \ldots, L_n^{N_n}\). Denote by \(\mathbb{L}(a)\) the sequence \(\{\mathcal{L}_n(a)\}_{n=1}^\infty\). Assume that subspaces \(L_n^j\) have the following property: for any probability measure \(w\) on \(X\) one has

\[
\|P_{L_n^j}^w\|_{L_\infty(w) \to L_\infty(w)} \leq K, \quad j \in [1, N_n], \quad n = 1, 2, \ldots.
\]

We note that as above

\[
d(f_\rho, L_n^j)_{L_2(\rho_X)} = \|f_\rho - P_{L_n^j}^\rho_X(f_\rho)\|_{L_2(\rho_X)}
\]

and by (4.12) and \(\|f_\rho\|_{L_\infty(\rho_X)} \leq M\) (we assume \(|y| \leq M\) a.e.) we get

\[
\|P_{L_n^j}^\rho_X(f_\rho)\|_{L_\infty(\rho_X)} \leq MK.
\]

Denote \(V_n^j := MKU(L_\infty(\rho_X)) \cap L_n^j\) and

\[
U_n := \bigcup_{j=1}^{N_n} V_n^j.
\]

Consider

\[
j(n) := \arg \min_{1 \leq j \leq N_n} d(f_\rho, L_n^j)_{L_2(\rho_X)}.
\]

Then

\[
f_{U_n} = f_{V_n^{j(n)}} = P_{L_n^{j(n)}}^\rho_X(f_\rho).
\]

For a given data \(z = \{(x_i, y_i)\}_{i=1}^m\) and a number \(n\) we define

\[
f_{z,n} := f_{z,U_n} = \arg \min_{f \in U_n} \mathcal{E}_z(f) = \arg \min_{1 \leq j \leq N_n} \min_{f \in V_n^j} \mathcal{E}_z(f).
\]

Denote by \(V_n := V_n^{j(z)}\) a set such that

\[
f_{z,U_n} = f_{z,V_n}.
\]

The following theorem is a nonlinear analog of Theorem 4.1.

**Theorem 4.4.** Let \(\rho\) be such that \(|y| \leq M\) a.e. Assume that \(\mathbb{L}(a)\) satisfies (4.12). For given \(m, r > 0, A \geq A_0(M, K, r, a)\) there exists an estimator \(f_z\) such that for any \(\rho\) satisfying

\[
\min_{1 \leq j \leq N_n} d(f_\rho, L_n^j)_{L_2(\rho_X)} \leq Dn^{-r}, \quad n = 1, 2, \ldots,
\]

we have

\[
\text{Prob}_{z \in Z^m}\{\|f_\rho - f_z\|_{L_2(\rho_X)} \leq C(D)A^{1/2}(\ln m/m)^{1/2+r} \} \\
\geq 1 - \exp\left(-c(M)A(m(\ln m)^{2r})^{1/2+r}\right).
\]

The proof of this theorem is close to the proof of Theorem 4.1 and the proof of Theorem 2.4 from [DKPT]. We will not present it here. We only point out that we set

\[
f_z := f_{z,U_n}
\]

with \(n := [(m/(A \ln m))^{1/2+r}] + 1\) and instead of Theorem 4.2 we use the following theorem from [DKPT] (see Theorem D).
Theorem 4.5. Let $\mathcal{H}$ be a compact subset of $L_\infty(\rho_X)$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \leq M$ a.e. Then, for all $\epsilon > 0$

$$\text{Prob}_{z \in Z^m} \{ \mathcal{E}(f_z, \mathcal{H}) - \mathcal{E}(f) \leq \epsilon \} \geq 1 - N(\mathcal{H}, \epsilon/(24M), L_\infty(\rho_X))2\exp\left(-\frac{m\epsilon}{C(M,B)}\right)$$

under assumption $\mathcal{E}(f) - \mathcal{E}(\tilde{f}) \leq B\epsilon$.

As an example of subspaces $L^j_n$ we may take the following subspaces of $L_\infty(w)$. Let $X$ be a compact subset of $\mathbb{R}^d$. Let $P_n$ denote the set of all partitions of $X$ into $n$ disjoint measurable (with regard to $w$) subsets. Let $p_j \in P_n$, $j = 1, \ldots, N_n$. Define $L^j_n$ as a subspace of all functions that are piecewise constant on the partition $p_j$. Then the property (4.1) is satisfied with $K = 1$. Therefore, we can use the results of this section for such approximation spaces.

5. Error estimates in the $L_p$ norm

In this section we obtain error estimates in the $L_p$-norm, $1 \leq p \leq \infty$. We assume that $\rho_X$ is the Lebesgue measure and $|y| \leq M$ a.e. We note that instead of assuming $\mu = 1$ in the arguments that follow it is sufficient to assume that $\mu \leq C$ with absolute constant $C$. Then we obtain the same results for $f_\mu$ instead of $f_\rho$. Let $\Omega$ be a bounded domain in $\mathbb{R}^d$. We assume for notational simplicity that the Lebesgue measure of $\Omega$ is 1 (otherwise we renormalize the Lebesgue measure). Let $\mathcal{K}_n(x, u)$ denote a continuous kernel defined on $\Omega \times \Omega$ with the following properties. Define

$$J_{\mathcal{K}_n}(f) := \int_\Omega f(u)\mathcal{K}_n(x, u)du.$$ 

Assume that the operator $J_{\mathcal{K}_n}$ is defined on the $L_\infty(\Omega)$ and rank($J_{\mathcal{K}_n}$) \leq n. Assume in addition that

(I) \quad \|J_{\mathcal{K}_n}\|_{L_\infty \to L_\infty} \leq K_1;

(II) \quad \|\mathcal{K}_n\|_{\infty} \leq K_2n;

and for any $x \in \Omega$

(III) \quad \int_\Omega |\mathcal{K}_n(x, u)|^2du \leq K_3n.

We define an estimator for $f_\rho$ by the formula:

$$f_z := \frac{1}{m} \sum_{i=1}^m y_i \mathcal{K}_n(x, x_i).$$
Then for the random variable $\xi(y, u) := yK_n(x, u)$ we obtain
\[
E(\xi) = \int_{\Omega} f_\rho(u)K_n(x, u)d\rho_x = \int_{\Omega} f_\rho(u)K_n(x, u)du = J_{K_n}(f_\rho).
\]
By property (III) we have for any $x \in \Omega$
\[
E(\xi^2) \leq M^2K_3n.
\]
Denote $K(n)$ the closure in $L_\infty$ of the range of the operator $J_{K_n}$. We note that for any $u$ we have $K_n(\cdot, u) \in K(n)$. We assume that for each $n$ there exists a set of points $\xi^1, \ldots, \xi^{N(n)} \in \Omega$ such that $N(n) \leq nK_4$ and for any $f \in K(n)$
\[
(IV) \quad \|f\|_\infty \leq K_5 \max_i |f(\xi^i)|.
\]
By Bernstein’s inequality (1.4) for each $\xi^i, l \in [1, N(n)]$ we have
\[
\text{Prob}_{z \in \mathbb{Z}^m}\{|J_{K_n}(f_\rho)(\xi^l) - f_z(\xi^l)| \geq \epsilon\} \leq 2\exp\left(-\frac{m\epsilon^2}{C(M, K_2, K_3)n}\right).
\]
Using (IV) we obtain
\[
(5.2) \quad \text{Prob}_{z \in \mathbb{Z}^m}\{|J_{K_n}(f_\rho) - f_z\|_\infty \leq K_5\epsilon\} \geq 1 - N(n)2\exp\left(-\frac{m\epsilon^2}{C(M, K_2, K_3)n}\right).
\]
We define the class $W_p^r(K, D)$ as the set of $f$ that satisfy the estimate:
\[
\|f - J_{K_n}(f)\|_p \leq Dn^{-r}, \quad n = 1, 2, \ldots, \quad 1 \leq p \leq \infty.
\]
Assume that $f_\rho \in W_p^r(K, D)$. We specify $\epsilon = A(ln m/m)^{1/r}$, $n = [\epsilon^{-1/r}] + 1$. Then (5.2) implies for $A \geq A_0(M, K_2, K_3, K_4)$
\[
\text{Prob}_{z \in \mathbb{Z}^m}\{|J_{K_n}(f_\rho) - f_z\|_\infty \leq K_5A(ln m/m)^{1/r}\} \geq 1 - w(m, A)
\]
and
\[
\text{Prob}_{z \in \mathbb{Z}^m}\{|f_\rho - f_z\|_p \leq (K_5 + D)A(ln m/m)^{1/r}\} \geq 1 - w(m, A)
\]
with $w(m, A) := \exp(-c(M, K_2, K_3)A^{2+1/r}ln m)$. We point out that we have obtained the $L_p$ estimates for $1 \leq p \leq \infty$. We formulate the result proved above as a theorem.

**Theorem 5.1.** Assume $f_\rho \in W_p^r(K, D)$ with some $1 \leq p \leq \infty$. Then the estimator $f_z$ defined by (5.1) with $n = [A^{-1/r}(m/(ln m))^{1/r}] + 1$ provides for $A \geq A_0(M, K_2, K_3, K_4)$
\[
\text{Prob}_{z \in \mathbb{Z}^m}\{|f_\rho - f_z\|_p \leq (K_5 + D)A(ln m/m)^{1/r}\} \geq 1 - \exp(-c(M, K_2, K_3)A^{2+1/r}ln m).
\]

We note that the estimator $f_z$ from Theorem 5.1 does not depend on $p$ and depends on $r$ (the choice of $n$ depends on $r$). We proceed to construction of an estimator that is universal for $r$. We denote
\[
W_p[K] := \{W_p^r(K, D)\}.
\]
**Theorem 5.2.** For a given collection $W_p[K]$ there exists an estimator $f_z$ such that if $f_\rho \in W_p^r(K, D)$ with some $r \leq R$ then for $A \geq A_0(M, K_2, K_3, K_4)$

$$\operatorname{Prob}_{z \in Z^m}\{\|f_\rho - f_z\|_p \leq C(R)(K_5 + D)A(\ln m/m)^{\frac{r}{1+2r}}\} \geq 1 - m^{-c(M, K, K_3)A^2}.$$ 

**Proof.** We define

$$A_0 := K_1; \quad A_s := K_2^s - K_{2s-1}, \quad s = 1, 2, \ldots; \quad A_s := J_{A_s}.$$ 

Therefore, for $s = 1, 2, \ldots$

$$A_s := J_{K_2^s - K_{2s-1}} = J_{K_2^s} - J_{K_{2s-1}}.$$ 

Using our assumption that $f_\rho \in W_p^r(K, D)$ we get for all $s$

$$\|A_s(f_\rho)\|_p \leq K2^{-rs}$$

with $K := (1 + 2^R)D$. We consider the following estimators

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

Similarly to (5.2) with $\epsilon = A((2^s/m) \ln m)^{1/2}$ we get for all $s \in [0, \log m]$

$$\|A_s(f_\rho) - f_{s,z}\|_\infty \leq K_5 A((2^s/m) \ln m)^{1/2}$$

with probability at least $1 - m^{-c(M, K_2, K_3)A^2}$, $A \geq A_0(M, K_2, K_3, K_4)$. We now consider only those $z$ that satisfy (5.4). We build an estimator $f_z$ on the base of the sequence $\{\|f_{s,z}\|_p\}_{s=0}^{[\log m]}$. First, if

$$\|f_{s,z}\|_p \leq (K_5A + K)((2^s/m) \ln m)^{1/2}, \quad s = 0, \ldots, [\log m],$$

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

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

Therefore, for $z$ satisfying (5.4) and (5.5) we get from (5.4)–(5.6), (5.3) that

$$\|f_\rho\|_p \leq C_1(R)(K_5 + D)A \sum_{s=0}^{\infty} \min(2^s \ln m/m)^{1/2}, 2^{-rs}) \leq C_2(R)(K_5 + D)A(\ln m/m)^{\frac{r}{1+2r}}.$$
Second, if (5.5) is not satisfied then we let \( l \in [0, \log m] \) be such that for \( s \in (l, \log m] \)

\[
\| f_{s,z} \|_p \leq (K_5 A + K)((2^s/m) \ln m)^{1/2}
\]

and

\[
\| f_{l,z} \|_p > (K_5 A + K)((2^l/m) \ln m)^{1/2}.
\]

We set \( n = 2^l \) and

\[
f_z := \frac{1}{m} \sum_{i=1}^{m} y_i K_{n}(x, x_i).
\]

Then by (5.4) we get from (5.8)

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

Therefore, by (5.3) with \( s = l \) we obtain

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

Let \( l_0 \) be such that

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

It is clear from the above two relations that \( l \leq l_0 \). Then for \( z \) satisfying (5.4) and not satisfying (5.5) we have

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

\[
\leq D 2^{-r l_0} + \sum_{s=l+1}^{l_0} (2K_5 A + K)((2^s/m) \ln m)^{1/2} + \sum_{s=0}^{l} K_5 A((2^s/m) \ln m)^{1/2}
\]

\[
\leq C(R)(K_5 + D) A(\ln m/m)^{1+r/2r}.
\]

Therefore, for \( z \) satisfying (5.4) we obtain

\[
\| f_{\rho} - f_z \|_p \leq C(R)(K_5 + D) A(\ln m/m)^{1+r/2r}.
\]

This completes the proof of Theorem 5.2.
THE ENTROPY IN THE LEARNING THEORY. ERROR ESTIMATES

References


