Mathematical Methods for Supervised Learning

Ronald DeVore, Gerard Kerkyacharian,
Dominique Picard, and Vladimir Temlyakov*

April 13, 2005

In honor of Steve Smale's 75-th birthday with the warmest regards of the authors

Abstract

Let ρ be an unknown Borel measure defined on the space $Z := X \times Y$ with $X \subset I\!R^d$ and $Y = [-M, M]$. Given a set $\mathbf{z}$ of m samples $z_i = (x_i, y_i)$ drawn according to ρ, the problem of estimating a regression function f_ρ using these samples is considered. The main focus is to understand what is the rate of approximation, measured either in expectation or probability, that can be obtained under a given prior $f_\rho \in \Theta$, i.e. under the assumption that f_ρ is in the set Θ, and what are possible algorithms for obtaining optimal or semi-optimal (up to logarithms) results. The optimal rate of decay in terms of m is established for many priors given either in terms of smoothness of f_ρ or its rate of approximation measured in one of several ways. This optimal rate is determined by two types of results. Upper bounds are established using various tools in approximation such as entropy, widths, and linear and nonlinear approximation. Lower bounds are proved using Kullback-Leibler information together with Fano inequalities and a certain type of entropy. A distinction is drawn between algorithms which employ knowledge of the prior in the construction of the estimator and those that do not. Algorithms of the second type which are universally optimal for a certain range of priors are given.

1 Introduction

We shall be interested in the problem of learning an unknown function defined on a set X which takes values in a set Y. We assume that X is a compact domain in $I\!R^d$ and $Y = [-M, M]$ is a finite interval in $I\!R$. The setting we adopt for this problem is called distribution free non-parametric estimation of regression. This problem has a long history in statistics and has recently drawn much attention in the work of Cucker and Smale [10] and amplified upon in Poggio and Smale [31]. We shall use the introduction to describe the setting and to explain our viewpoint of this problem which is firmly oriented

*This research was supported by the Office of Naval Resarch Contracts ONR-N00014-03-1-0051, ONR/DEPSCoR N00014-03-1-0675 and ONR/DEPSCoR N00014-00-1-0470; the Army Research Office Contract DAAD 19-02-1-0028; the AFOSR Contract UF/USAF F49620-03-1-0381; and NSF contracts DMS-0221642 and DMS-0200187

in approximation theory. Later in this introduction, we shall explain the new results obtained in this paper. We have written the paper to be as self contained as possible and accessible to researchers in various disciplines. As such, parts of the paper may seem pedestrian to some researchers but we hope that they will find other aspects of the paper to be of interest.

1.1 The learning problem

There are many examples of learning problems given in [31]. We shall first describe one such problem whose sole purpose is to aid the reader to understand the setting and the assumptions we put forward. Consider the problem of a bank wanting to decide whether or not to give an individual a loan. The bank will ask the potential client to answer several questions which are deemed to be related to how he will perform in paying back the loan. Sample questions could be age, income, marital status, credit history, home ownership, amount of the loan, etc. The answer to these questions form a point in $\mathbb{R}^d$, where d is the number of questions. We assume that d is fixed and each potential client is asked the same questions. The bank will have a data set (history) of past customers and how they have performed in paying back their loans. We denote by y the profit (or loss if negative) the bank has made on a particular loan. Thus a point $z := (x, y) \in Z := X \times Y$ represents a (potential) client's answers (x) and the (potential) profit y the bank has made (or would make) on the loan. The data collection will be denoted by $\mathbf{z}$ and consists of points $(x_i, y_i) \in \mathbb{R}^{d+1}$ where x_i is the answers given by the i-th customer and y_i is the profit or loss the bank made from that loan.

Notice there are two distributions lurking in the background of this problem. The first is the distribution of answers $x \in X$. Typically, several potential customers would have the same answers x and some x are more likely than others. So our first distribution is on X. The second distribution relates to the profit (y) the bank will make on the loan. Given an x there will be several different customers with these same answers and therefore there will be several different values y associated to this x. Thus sitting over x there is a probability distribution in Y. The bank is interested in learning the function f defined on X which describes the expected profit $f(x)$ over the collection of all potential customers with answers x. It is this function f that we wish to learn. What we have available are the past records of loans. This corresponds to the set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ which is a subset of Z^m. This set incorporates all of the information we have about the two unknown distributions. Our problem then is to estimate f by a function $f_{\mathbf{z}}$ determined in some way from the set $\mathbf{z}$.

A precise mathematical formulation of this type of problem (see [19] or [10]) incorporates both probability distributions into one (unknown) Borel probability measure ρ defined on $Z = X \times Y$. The conditional probability measure $\rho(y|x)$ represents the probability of an outcome y given the data x. The marginal probability measure ρ_X defined for $S \subset X$ by $\rho_X(S) = \rho(S \times Y)$ describes the distribution on X. The function f_ρ we are trying to learn is then

$$f_\rho(x) := \int_Y y d\rho(y|x). \tag{1.1}$$

The function f_ρ is known in statistics as the *regression function* of ρ. One should note that f_ρ is the minimizer of

$$\mathcal{E}(f) := \mathcal{E}_\rho(f) := \int_Z (f(x) - y)^2 d\rho \tag{1.2}$$

among all functions $f : X \to Y$. This formula will motivate some of the approaches to constructing an $f_{\mathbf{z}}$.

Notice that if we put $\epsilon = Y - f(X)$, then we are not assuming that ϵ and X are independent although there is a large body of statistical literature which makes this assumption. While these theories do not directly apply to our setting, they utilize several of the same techniques we shall encounter such as the utilization of entropy and the construction of estimators through minimal risk.

There may be other settings in which f_ρ is not the function we wish to learn. For example, if we replace the L_2 norm in (1.2) by the L_1 norm then in place of the mean $f_\rho(x)$ we would want to learn the median of $y|x$. In this paper, we shall be interested in learning f_ρ or some variant of this function.

Our problem then is given the data $\mathbf{z}$, how to find a good approximation $f_{\mathbf{z}}$ to f_ρ. We shall call a mapping $I\!E_m$ that associates to each $\mathbf{z} \in Z^m$ a function $f_{\mathbf{z}}$ defined on X to be an *estimator*. By an *algorithm*, we shall mean a family of estimators $\{I\!E_m\}_{m=1}^\infty$. To evaluate the performance of estimators or algorithms, we must first decide how to measure the error in the approximation of f_ρ by $f_{\mathbf{z}}$. The typical candidates to measure error are the $L_p(X, \rho_X)$ norms [1]:

$$\|g\|_{L_p(X,\rho_X)} := \begin{cases} \left(\int_X |g(x)|^p d\rho_X \right)^{1/p}, & 1 \le p < \infty, \\ esssup_{x \in X} |g(x)|, & p = \infty. \end{cases} \tag{1.3}$$

Other standard choices in statistical litterature correspond to taking measures other than ρ_X in the L_p norm, for instance the Lebesgue measure. In this paper, we shall have as our goal to obtain approximations to f_ρ with the error measured in the $L_2(X, \rho_X)$ norm. However, as we shall see estimates in the $\mathcal{C}(X)$ norm [2] are an important tool in such an analysis.

Having set our problem, what kind of estimators $f_{\mathbf{z}}$ could we possibly construct? The most natural approach (and the one most often used in statistics) is to choose a class of functions $\mathcal{H}$ which is to be used in the approximation, i.e. $f_{\mathbf{z}}$ will come from the class $\mathcal{H}$. This class of functions is called the *hypothesis space* in learning theory. A typical choice for $\mathcal{H}$ is a ball in a linear space of finite dimension n or in a nonlinear manifold of dimension n. (The best choice for the dimension n (depending on m) will be a critical issue which will emerge in the analysis.) For example, in the linear case, we might choose a space of polynomials, or splines, or wavelets, or radial basis functions. Candidates in the nonlinear case could be free knot splines or piecewise polynomials on adaptively generated

[1] The space $L_p = L_p(X)$ will always be understood to be with respect to Lesbesgue measure. Spaces with respect to other measures will always have further amplification such as $L_p(X, \rho_X)$

[2] Here and later $\mathcal{C}(X)$ denotes the space of continuous functions defined on X

partitions or n-term approximation from a basis of wavelets, or radial basis functions, or ridge functions (a special case of this would correspond to neural networks). Once this choice is made, the problem is then given $\mathbf{z}$ how do we find a good approximation $f_{\mathbf{z}}$ to f_ρ from $\mathcal{H}$. We will turn to that problem in a moment but first we want to discuss how to measure the performance of such an approximation scheme.

As mentioned earlier, we shall primarily measure the approximation error in the $L_2(X, \rho_X)$ norm. If we have a particular approximant $f_{\mathbf{z}}$ to f_ρ in hand, the quality of its performance is measured by

$$\|f_\rho - f_{\mathbf{z}}\|. \tag{1.4}$$

Throughout the paper, we shall use the default notation $\|g\| := \|g\|_{L_2(X,\rho_X)}$. Other norms will have an apropriate subscript. The error (1.4) clearly depends on $\mathbf{z}$ and therefore has a stochastic nature. As a result, it is generally not possible to say anything about (1.4) for a fixed $\mathbf{z}$. Instead, we can look at behavior in probability as measured by

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| > \eta\}, \quad \eta > 0 \tag{1.5}$$

or the expected error

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) = \int_{Z^m} \|f_\rho - f_{\mathbf{z}}\| d\rho^m, \tag{1.6}$$

where the expectation is taken over all realizations $\mathbf{z}$ obtained for a fixed m and ρ^m is the m-fold tensor product of ρ.

If we have done things correctly, this expected error should tend to zero as $m \to \infty$ (the law of large numbers). How fast it tends to zero depends on at least three things: (i) the nature of f_ρ, (ii) the approximation properties of the space $\mathcal{H}$, (iii) how well we did in constructing the estimators $f_{\mathbf{z}}$. We shall discuss each of these components subsequently.

The probability

$$\rho^m\{\mathbf{z} : \|f_z - f_\rho\| > \eta\} \tag{1.7}$$

measures the confidence we have that the estimator is accurate to tolerance η. We are interested in the decay of (1.7) as $m \to \infty$ and η increases.

Notice that we really do not know the norm $\|\cdot\|$ because we do not know the measure ρ. This does not prevent us from formulating theorems in this norm however. An important observation is that for any probability measure ρ_X, we have

$$\|f\|_{L_2(X,\rho_X)} \le \|f\|_{\mathcal{C}(X)}. \tag{1.8}$$

Thus, bounds on the goodness of fit in $\mathcal{C}(X)$ imply the same bounds in $L_2(X, \rho_X)$. While obtaining estimates through $\mathcal{C}(X)$ provides a quick fix to not knowing ρ, it may be a nonoptimal approach.

1.2 The role of approximation theory

The expected error (1.6) has two components which are standard in statistics. One is how well we can approximate f_ρ by the elements of $\mathcal{H}$ (called the bias) and the second is the stochastic nature of $\mathbf{z}$ (the variance). We discuss the first of these now and show how it

influences the form of the results we can expect. We phrase our discussion in the context of approximation in a general Banach space $\mathcal{B}$ even though our main interest will be the case $\mathcal{B} = L_2(X, \rho_X)$. Understanding how well $\mathcal{H}$ does in approximating functions is critical to understanding the advantages and disadvantages of such a choice. The performance of approximation by the elements of $\mathcal{H}$ is the subject of approximation theory. This subject has a long and important history which we cannot give in its entirety. Rather we will give a coarse resolution of approximation theory in order to not inundate the reader with a myriad of results that are difficult to absorb on first exposure. We will return to this subject again in more detail in §2.3.

Given a set $\mathcal{H} \subset \mathcal{B}$, and a function $f \in \mathcal{B}$, we define

$$\operatorname{dist}(f, \mathcal{H})_{\mathcal{B}} := \inf_{S \in \mathcal{H}} \|f - S\|_{\mathcal{B}}. \tag{1.9}$$

More generally, for any compact set $K \subset \mathcal{B}$, we define

$$\operatorname{dist}(K, \mathcal{H})_{\mathcal{B}} := \sup_{f \in K} \operatorname{dist}(f, \mathcal{H})_{\mathcal{B}}. \tag{1.10}$$

Certainly, $f_{\mathbf{z}}$ will never approximate f_ρ (in the $\mathcal{B}$ sense) with error better than (1.9). However, in general it will do (much) worse for two reasons. The first is that we only have (partial) information about f_ρ from the data $\mathbf{z}$. The second is that the data $\mathbf{z}$ is noisy in the sense that for each x the value $y|x$ is stochastic.

Approximation theory seeks quantitative descriptions of approximation given by sequence of spaces $\mathcal{S}_n$, $n = 1, 2, \ldots$, which will be used in the approximation. The spaces could be linear of dimension n or nonlinear depending on n parameters. A typical result is that given a compact set $K \subset \mathcal{B}$, approximation theory determines the best exponent $r = r(K) > 0$ [3] for which

$$\operatorname{dist}(K, \mathcal{S}_n)_{\mathcal{B}} \le C_K n^{-r}, \quad n = 1, 2, \ldots. \tag{1.11}$$

Such results are known in all classical settings in which $\mathcal{B}$ is one of the L_p spaces (with respect to Lebesgue measure) and K is given by a smoothness condition. Sometimes it is even possible to describe the functions which are approximated with a specified approximation order. The approximation class $\mathcal{A}^r := \mathcal{A}^r((\mathcal{S}_n), \mathcal{B})$ consists of all functions f such that

$$\operatorname{dist}(f, \mathcal{S}_n)_{\mathcal{B}} \le M n^{-r}, \quad n = 1, 2, \ldots. \tag{1.12}$$

The smallest $M = M(f)$ for which (1.12) is valid is by definition the semi-norm $|f|_{\mathcal{A}^r}$ in this space.

To orient the reader let us give a classical example in approximation theory in which we approximate continuous functions f in the $\mathcal{C}(X)$ norm (i.e the uniform norm on X). For simplicity, we take X to be $[0, 1]$. We consider first the case when $\mathcal{S}_n$ is the linear n-dimensional space consisting of all piecewise constant functions on the uniform partition of X into n disjoint intervals. (Notice that the approximating functions are not continuous.) In this case the space $\mathcal{A}^r$, for $0 < r \le 1$, is precisely the Lipshitz space $Lip\, r$ [4] (see e.g.

[3] We shall exclusively use the parameter r to denote a rate of approximation in this paper.

[4] If the reader is unfamiliar with the space $Lip\, r$ then he may wish to look forward to §2 where we give a general discussion of smoothness spaces

[16]). The semi-norm for $\mathcal{A}^r$ is equivalent to the above $Lip\, r$ semi-norm. In other words, we can get an approximation rate $\operatorname{dist}(f, \mathcal{S}_n)_{\mathcal{C}(X)} = O(n^{-r})$ if and only if $f \in Lip\, r$.

Let us consider a second related example of nonlinear approximation. Here we again approximate in the norm $\mathcal{C}(X)$ by piecewise constants but allow the partition of $[0,1]$ to be arbitrary except that the number of intervals is again restricted to be n. The corresponding space $\mathcal{S}_n$ is now a nonlinear manifold which is described by $2n-1$ parameters (the $n-1$ breakpoints and the n constant values on the intervals). In this case, the approximation classes are again known (see [16]), but we mention only the case $r = 1$. In this case, $\mathcal{A}^1 = BV \cap \mathcal{C}(X)$, where BV is the space of functions of bounded variation on $[0,1]$. Here we can see the distinction between linear and nonlinear approximation. The class BV is much larger than $Lip\ 1$. So we obtain the performance $O(n^{-1})$ for a much larger class in the nonlinear case. Note however that if $f \in Lip\ 1$, then the nonlinear method does not improve the approximation rate; it is still $O(n^{-1})$. On the other hand, for general functions in $BV \cap C$, we can say nothing at all about the linear approximation rate while the nonlinear rate is $O(n^{-1})$.

The problem with using these approximation results directly in our learning setting is that we do not know the function f_ρ. Nevertheless, a large portion of statistics and learning theory proceeds under the assumption that f_ρ is in a known set Θ. Such assumptions are known as priors in statistics. We shall denote such priors by $f_\rho \in \Theta$. Typical choices of Θ are compact sets determined by some smoothness condition or by some prescribed rate of decay for a specific approximation process. We shall denote generic smoothness spaces by W. Given a normed (or quasi-normed) space $\mathcal{B}$, we denote its unit ball by $u(\mathcal{B})$. We denote a ball of radius R different from one by $b_R(\mathcal{B})$. [5] If we do not wish to specify the radius we simply write $b(\mathcal{B})$.

If we assume that f_ρ is in some known compact set K and nothing more, then the best estimate we can give for the bias term is

$$\operatorname{dist}(f_\rho, \mathcal{H})_{\mathcal{B}} \le \operatorname{dist}(K, \mathcal{H})_{\mathcal{B}}. \tag{1.13}$$

The question becomes what is a good set $\mathcal{H}$ to use in approximating the elements of K. These questions are answered by concepts in approximation theory known as widths or entropy numbers as we shall now describe.

Suppose that we decide to use linear spaces in our construction of $f_{\mathbf{z}}$. We might then ask what is the best linear space to choose. The vehicle for making this decision is the concept of Kolmogorov widths. Given a centrally symmetric compact set K from a Banach space $\mathcal{B}$, the Kolmogorov n-width is defined by

$$d_n(K, \mathcal{B}) := \inf_{\mathcal{L}_n} \operatorname{dist}(K, \mathcal{L}_n)_{\mathcal{B}} \tag{1.14}$$

where $\inf_{\mathcal{L}_n}$ is taken over all n-dimensional linear subspaces $\mathcal{L}_n$ of $\mathcal{B}$. In other words, the Kolmogorov n-width gives the best possible error in approximating K by n-dimensional linear subspaces. Thus, the best choice of $\mathcal{L}_n$ (from the viewpoint of approximation theory) is to choose $\mathcal{L}_n$ as a space that gives (or nearly gives) the infimum in (1.14). It is usually impossible to find the *best* n-dimensional approximating subspace for K and we

[5] We use lower case b for balls in order to not have confusion with the other uses of B in this paper.

have to be satisfied with a *near optimal* sequence $(\mathcal{L}_n)$ of subspaces by which we mean

$$\operatorname{dist}(K, \mathcal{L}_n) \le C d_n(K, \mathcal{B}), \quad n = 1, 2, \dots, \tag{1.15}$$

with C an absolute constant.

For bounded sets in any of the classical smoothness spaces W and for approximation in $\mathcal{B} = L_p$ (with Lebesgue measure), the order of decay of the n-widths is known. However, we should caution that in some of the deeper theorems, (near) optimizing spaces are not known explicitly. As an example, for any ball in one of the Lipschitz spaces $Lip\, s$, $0 < s \le 1$, introduced above, the n-width is known to behave like $O(n^{-s})$ and therefore piecewise constants on a uniform partition form a sequence of near optimal linear subspaces. There is a similar concept of nonlinear widths (see [1, 14]) to describe best n-dimensional manifolds for nonlinear approximation. We give one formulation of nonlinear widths in §4.2

Another way of measuring the approximability of a set is through covering numbers. Given a compact set K in a Banach space $\mathcal{B}$, for each $\epsilon > 0$ the *covering number* $N(\epsilon, K)_{\mathcal{B}}$ is the smallest number of balls in $\mathcal{B}$ of radius ϵ which cover K. We shall use the default notation

$$N(\epsilon, K) = N(\epsilon, K)_{\mathcal{C}(X)} \tag{1.16}$$

for the covering numbers in $\mathcal{C}(X)$. The logarithm

$$H(\epsilon, K) := H(\epsilon, K)_{\mathcal{B}} := \log_2 N(\epsilon, K)_{\mathcal{B}} \tag{1.17}$$

of the covering number is the *Kolmogorov entropy* of K in $\mathcal{B}$. From the Kolmogorov entropy we obtain the entropy numbers of K defined by

$$\epsilon_n(K) := \epsilon_n(K, \mathcal{B}) := \inf\{\epsilon : H(\epsilon, K)_{\mathcal{B}} \le n\}. \tag{1.18}$$

The entropy numbers are very closely related to nonlinear widths. For example, if $\mathcal{B}$ is chosen as any of the L_p spaces, $1 \le p \le \infty$, and K is a unit ball of an isotropic smoothness space (Besov or Sobolev) which is compactly embedded in $\mathcal{B}$, then the nonlinear width of K decays like $O(n^{-r})$ if and only if $\epsilon_n(K) = O(n^{-r})$. Moreover, one can obtain this approximation rate through a simple nonlinear approximation method such as either wavelet thresholding or piecewise polynomial approximation on adaptively generated partitions (see [13]).

1.3 Measuring the quality of the approximation

We have already discussed possible norms to measure how well $f_{\mathbf{z}}$ approximates f_ρ. We shall almost always use the $L_2(X, \rho_X)$ norm and it is our default norm (denoted simply by $\|\cdot\|$). Given this norm one then considers the expected error (1.6) as a measure of how well the $f_{\mathbf{z}}$ approximates f_ρ. We have also mentioned measuring accuracy in probability. Given a bound for $\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| > \eta\}$, we can obtain a bound for the expected error from

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) = \int_0^\infty \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| > \eta\} d\eta. \tag{1.19}$$

Bounding probabilities like ρ^m utilizes concentration of measure inequalities. Let ρ be a Borel probability measure on $Z = X \times Y$. If ξ is a random variable (a real valued function on Z) then

$$E(\xi) := \int_Z \xi d\rho; \quad \sigma^2(\xi) := \int_Z (\xi - E(\xi))^2 d\rho \tag{1.20}$$

are its expectation and variance respectively. The law of large numbers says that drawing samples $\mathbf{z}$ from Z, the sum $\frac{1}{m}\sum_{i=1}^m \xi(z_i)$ will converge to $E(\xi)$ with high probability as $m \to \infty$. There are various quantitative versions of this, known as concentration of measure inequalities. We mention one particular inequality (known as Bernstein's inequality) which we shall employ in the sequel. This inequality says that if $|\xi(z) - E(\xi)| \leq M_0$ a.e. on Z, then for any $\eta > 0$

$$\rho^m\{\mathbf{z} \in Z^m : |\frac{1}{m}\sum_{i=1}^m \xi(z_i) - E(\xi)| \geq \eta\} \leq 2\exp(-\frac{m\eta^2}{2(\sigma^2(\xi) + M_0\eta/3)}). \tag{1.21}$$

1.4 Constructing estimators: empirical risk minimization

Suppose that we have decided on a set $\mathcal{H}$ which we shall use in approximating f_ρ, i.e. $f_{\mathbf{z}}$ should come from $\mathcal{H}$. We need still to address the question of how to find an estimator $f_{\mathbf{z}}$ to f_ρ. We shall use empirical risk minimization (least squares data fitting). This is of course a widely studied method in statistics. This subsection describes this method and introduces some fundamental concepts as presented in Cucker and Smale [10].

Empirical risk minimization is motivated by the fact that f_ρ is the minimizer of

$$\mathcal{E}(f) := \mathcal{E}_\rho(f) := \int_Z (f(x) - y)^2 d\rho. \tag{1.22}$$

That is (see [2]),

$$\mathcal{E}(f_\rho) = \inf_{f \in L_2(X,\rho_X)} \mathcal{E}(f). \tag{1.23}$$

Notice that for any $f \in L_2(X, \rho_X)$, we have

$$\begin{aligned} \mathcal{E}(f) - \mathcal{E}(f_\rho) &= \int_Z \{(y-f)^2 - (y-f_\rho)^2\} d\rho = \int_Z \{f^2 - 2y(f - f_\rho) - f_\rho^2\} d\rho \\ &= \int_X \{f^2 - 2f_\rho f + f_\rho^2\}\, d\rho_X = \|f - f_\rho\|^2. \end{aligned} \tag{1.24}$$

We use this formula frequently when we try to assess how well a function f approximates f_ρ.

Properties (1.22) and (1.23) suggest to consider the problem of minimizing the empirical variance

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 \tag{1.25}$$

over all $f \in \mathcal{H}$. We denote by

$$f_{\mathbf{z}} := f_{\mathbf{z},\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f), \tag{1.26}$$

the so-called *empirical minimizer*. We shall use this approach frequently in trying to find an approximation $f_{\mathbf{z}}$ to f_ρ. Given a finite ball in a linear or nonlinear finite dimensional space, the problem of finding $f_{\mathbf{z}}$ is numerically executable.

We turn now to the question of estimating $\|f_\rho - f_{\mathbf{z}}\|$ under this choice of $f_{\mathbf{z}}$. There is a long history in statistics of using entropy of the set $\mathcal{H}$ in bounding this error in one form or another. We shall present the core estimate of Cucker and Smale [10] which we shall employ often in this paper.

Our first observations center around the minimizer

$$f_{\mathcal{H}} := \arg\min_{f\in\mathcal{H}} \mathcal{E}(f). \tag{1.27}$$

From (1.24), it follows that $f_{\mathcal{H}}$ is the best approximation to f_ρ from $\mathcal{H}$:

$$\|f_\rho - f_{\mathcal{H}}\| = dist(f, \mathcal{H}). \tag{1.28}$$

If $\mathcal{H}$ were a linear space then $f_{\mathcal{H}}$ is unique and $f_\rho - f_{\mathcal{H}}$ is orthogonal to $\mathcal{H}$. We shall typically work with bounded sets $\mathcal{H}$ and so this kind of orthogonality needs more care. Suppose that $\mathcal{H}$ is any closed convex set. Then 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,

$$0 \le \|f_\rho - f_{\mathcal{H}} - \epsilon g\|^2 - \|f_\rho - f_{\mathcal{H}}\|^2 = -2\epsilon \int_X (f_\rho - f_{\mathcal{H}}) g \, d\rho_X + \epsilon^2 \int_X g^2 \, d\rho_X. \tag{1.29}$$

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

$$\int_X (f_\rho - f_{\mathcal{H}})(f - f_{\mathcal{H}}) \, d\rho_X \le 0, \quad f \in \mathcal{H}. \tag{1.30}$$

Then letting $\epsilon = 1$ we see that $\|f_\rho - f\| > \|f_\rho - f_{\mathcal{H}}\|$ whenever $f \ne f_{\mathcal{H}}$ and so $f_{\mathcal{H}}$ is unique. Also, (1.30) gives

$$\|f_{\mathcal{H}} - f_{\mathbf{z}}\|^2 \le \|f_\rho - f_{\mathbf{z}}\|^2 - \|f_\rho - f_{\mathcal{H}}\|^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}). \tag{1.31}$$

Of course, we cannot find $f_{\mathcal{H}}$ but it is useful to view it as our target in the construction of the $f_{\mathbf{z}}$.

We are left with understanding how well $f_{\mathbf{z}}$ approximates $f_{\mathcal{H}}$ or said in another way how the empirical minimization compares to the actual minimization (1.27). For $f : X \to Y$, the *defect function*

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

measures the difference between the true and empirical variances. Since $\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \ge \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$, returning to (1.31), we find

$$\|f_{\mathcal{H}} - f_{\mathbf{z}}\|^2 \le L_{\mathbf{z}}(f_{\mathbf{z}}) - L_{\mathbf{z}}(f_{\mathcal{H}}). \tag{1.32}$$

The approach to bounding quantities like $L_{\mathbf{z}}(f)$ is to use Bernstein's inequality. If the random variable $(y - f(x))^2$ satisfies $|y - f(x)| \leq M_0$ for $x, y \in Z$, then $\sigma^2 := \sigma^2((y - f(x))^2) \leq M_0^4$ and Bernstein's inequality gives

$$\rho^m\{\mathbf{z} \in Z^m : |L_{\mathbf{z}}(f)| \geq \eta\} \leq 2\exp(-\frac{m\eta^2}{2(M_0^4 + M_0^2\eta/3)}), \quad \eta > 0. \tag{1.33}$$

The estimate (1.33) suffices to give a bound for the second term $L_{\mathbf{z}}(f_{\mathcal{H}})$ in (1.32). However, it is not sufficient to bound the first term because the function $f_{\mathbf{z}}$ is changing with $\mathbf{z}$. Cucker and Smale utilize covering numbers to bound $L_{\mathbf{z}}(f_{\mathbf{z}})$ as follows. They assume that $\mathcal{H}$ is compact in $\mathcal{C}(X)$. Then, under the assumption

$$|y - f(x)| \leq M_0, \quad (x, y) \in Z, \ f \in \mathcal{H}, \tag{1.34}$$

it is shown in [10] (see Theorem B) that

$$\rho^m\{\mathbf{z} : \sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| \geq \eta\} \leq N(\eta/(8M_0), \mathcal{H})_{\mathcal{C}(X)} \exp(-\frac{m\eta^2}{4(2\sigma_0^2 + M_0^2\eta/3)}), \quad \eta > 0, \tag{1.35}$$

where $\sigma_0^2 := \sup_{f\in\mathcal{H}} \sigma^2((y - f(x))^2)$. Note again that from (1.34) we derive $\sigma_0^2 \leq M_0^4$.

Putting all of this together (see [10] for details), one obtains the following theorem:

Theorem C [[10]: *Let $\mathcal{H}$ be a compact subset in $\mathcal{C}(X)$. If (1.34) holds, then, for all $\eta > 0$,*

$$\rho^m\{\mathbf{z} \in Z^m : \|f_{\mathcal{H}} - f_{\mathbf{z},\mathcal{H}}\|^2 \geq \eta\} \leq 2N(\eta/(16M_0), \mathcal{H})_{\mathcal{C}(X)} \exp(-\frac{m\eta^2}{8(4\sigma_0^2 + M_0^2\eta/3)}), \tag{1.36}$$

where $\sigma_0^2 := \sup_{f\in\mathcal{H}} \sigma^2((f(x) - y)^2)$.

A second technique of Cucker and Smale gives an improved estimate to (1.36). This second approach makes the stronger assumption that either $f_\rho \in \mathcal{H}$ or the minimizer $f_{\mathcal{H}}$ and all of the estimators $f_{\mathbf{z}}$ come from a set $\mathcal{H}$ which is not only compact but also convex in $\mathcal{C}(X)$.

Theorem C* [10] *Let $\mathcal{H}$ be either a compact and convex subset of $\mathcal{C}(X)$ or a compact subset of $\mathcal{C}(X)$ for which $f_\rho \in \mathcal{H}$. If (1.34) holds, then, for all $\eta > 0$*

$$\rho^m\{\mathbf{z} \in Z^m : \|f_{\mathcal{H}} - f_{\mathbf{z},\mathcal{H}}\|^2 \geq \eta\} \leq 2N(\eta/(24M_0), \mathcal{H}) \exp(-\frac{m\eta}{288M_0^2}). \tag{1.37}$$

There is a long history in statistics of obtaining bounds, like those given above, through entropy and concentration of measure inequalities. It would be impossible for us to give proper credit here to all of the relevant works. However, a good start would be to look at the books of S. Van de Geer [38] and L. Györfi, M. Kohler, A. Krzyzak, and H. Walk [19] and the references therein. In this paper, we will, partly for the sake of simplicity, restrict the exposition to concentration bounds related to Bernstein's inequality. However, many refinements, in particular functional ones, can be found in the probability literature and used with profit (see e.g. Ledoux and Talagrand [28] and Talagrand [34]).

1.5 Approximating f_ρ: first bounds for error

For the remainder of this paper, we shall limit ourselves to the following setting. We assume that X is a bounded set in $\mathbb{R}^d$ which we can always take to be a cube. We also assume as before that Y is contained in the interval $[-M, M]$. It follows that f_ρ is bounded: $|f_\rho(x)| \le M$, $x \in X$.

Let us return to the estimates of the previous section. We know that $f_{\mathcal{H}}$ is the best approximation from $\mathcal{H}$ to f_ρ in $L_2(X, \rho_X)$ and so the bias term satisfies

$$\|f_\rho - f_{\mathcal{H}}\| = \operatorname{dist}(f_\rho, \mathcal{H})_{L_2(X,\rho_X)} =: \operatorname{dist}(f_\rho, \mathcal{H}). \tag{1.38}$$

To apply Theorem C, we need to know that

$$|f(x) - y| \le M_0, \quad (x, y) \in Z, \ f \in \mathcal{H}. \tag{1.39}$$

If this is the case then we have for any $\eta > 0$,

$$\rho^m\{\mathbf{z} \in Z^m : \|f_{\mathbf{z}} - f_{\mathcal{H}}\| \ge \eta\} \le 2N(\eta^2/(8M_0), \mathcal{H}) e^{-\frac{m\eta^4}{8(4\sigma_0^2 + M_0^2\eta^2/3)}}. \tag{1.40}$$

This gives that for any $\eta > 0$

$$\|f_\rho - f_{\mathbf{z}}\| \le \operatorname{dist}(f_\rho, \mathcal{H}) + \eta, \quad \mathbf{z} \in \Lambda_m(\eta), \tag{1.41}$$

for a set $\Lambda_m(\eta)$ which satisfies

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le 2N(\eta^2/(8M_0), \mathcal{H}) e^{-\frac{m\eta^4}{8(4\sigma_0^2 + M_0^2\eta^2/3)}}. \tag{1.42}$$

Since $\sigma_0^2 \le M_0^2$, this last estimate can be restated as

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le 2Ne^{-c_1 m\eta^4}, \quad \eta > 0, \tag{1.43}$$

with $N := N(\eta^2/(8M_0), \mathcal{H})$ and $c_1 := [32M_0^2(1 + M_0^2/3)]^{-1}$. Indeed, if $\eta > 2M_0$, then from (1.39) we conclude $\|f_{\mathcal{H}} - f_{\mathbf{z}}\| \le \eta$ for all $\mathbf{z} \in Z^m$, so that (1.43) trivially holds. On the other hand, if $\eta \le 2M_0$ then the denominator in the exponential (1.42) is $\le 32M_0^2(1 + M_0^2/3)$.

If we do a similar analysis using Theorem C* in place of Theorem C, we derive that

$$\|f_\rho - f_{\mathbf{z}}\| \le \operatorname{dist}(f_\rho, \mathcal{H}) + \eta, \quad \mathbf{z} \in \Lambda_m(\eta), \tag{1.44}$$

where

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le 2Ne^{-c_2 m\eta^2}. \tag{1.45}$$

The game is now clear. Given m, we need to choose the set $\mathcal{H}$. This set will typically depend on m. The question is what is a good choice for $\mathcal{H}$ and what type of estimates can be derived from (1.41) for this choice. Notice the two competing issues. We would like $\mathcal{H}$ to be large in order that the bias term $\operatorname{dist}(f_\rho, \mathcal{H})$ is small. On the other hand, we would like to keep $\mathcal{H}$ small so that its covering numbers $N(\eta^2/(8M_0), \mathcal{H})$ are small. This is a common situation in statistical estimation, leading to the desire to balance the bias and variance terms.

Cucker and Smale [10] mention two possible settings in which to apply Theorems C and C*. We want to carry their line of reasoning a little further to see what this gives for the actual approximation error. In the first setting, we assume that Θ is a compact subset of $\mathcal{C}(X)$ and therefore Θ is contained in a finite ball in $\mathcal{C}(X)$. Given m, we choose $\mathcal{H} = \Theta$. This means that (1.39) will be satisfied for some M_0. Since Θ is compact in $\mathcal{C}(X)$, its entropy numbers $\epsilon_n(\Theta)$ tend to zero with $n \to \infty$. If these entropy numbers behave like

$$\epsilon_n(\Theta) \le Cn^{-r}, \tag{1.46}$$

then $N(\eta, \Theta) \le e^{c_0 \eta^{-1/r}}$ and $\mathrm{dist}(f_\rho, \Theta) = 0$, and (1.41) gives

$$\|f_\rho - f_{\mathbf{z}}\| \le \eta, \quad \mathbf{z} \in \Lambda_m(\eta), \tag{1.47}$$

where

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le e^{c_0 \eta^{-2/r} - c_1 m \eta^4}. \tag{1.48}$$

In other words, for any $\eta > 0$, we have

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le e^{c_0 \eta^{-2/r} - c_1 m \eta^4}. \tag{1.49}$$

The critical value of η occurs when $c_1 m \eta^4 = c_0 \eta^{-2/r}$, i.e. for $\eta = \eta_m = (\frac{c_0}{c_1 m})^{\frac{r}{4r+2}}$ and we obtain

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C\{ \begin{array}{ll} e^{-cm\eta^4}, & \eta \ge 2\eta_m, \\ 1, & \eta \le 2\eta_m, \end{array} \tag{1.50}$$

in particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le Cm^{-\frac{r}{4r+2}}. \tag{1.51}$$

This situation is improved if we use Theorem C* in place of Theorem C in the above analysis. This allows us to replace $e^{-c_1 m \eta^4}$ by $e^{-c_2 m \eta^2}$ in the above estimates and now the critical value of η is $\eta_m^* = cm^{-\frac{r}{2r+2}}$ and we obtain the following Corollary.

Corollary 1.1 *Let Θ be either a compact subset of $\mathcal{C}(X)$ or a compact subset of $\mathcal{C}(X)$ for wich $f_\rho \in \Theta$ and*

$$\epsilon_n(\Theta) \le Cn^{-r}, \quad n = 1, 2 \ldots. \tag{1.52}$$

Then, by taking $\mathcal{H} := \Theta$, we obtain the estimate for $m = 1, 2, \ldots$,

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z},\Theta}\| \ge \eta\} \le C\{ \begin{array}{ll} e^{-cm\eta^2}, & \eta \ge cm^{-\frac{r}{2r+2}}, \\ 1, & \eta \le cm^{-\frac{r}{2r+2}}, \end{array} \tag{1.53}$$

In particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z},\Theta}\|) \le Cm^{-\frac{r}{2r+2}}. \tag{1.54}$$

Example: The simplest example of a prior Θ which satisfies the asumptions of the corollary is $\Theta := b(W)$ where W is the Sobolev space $W^s(L_\infty(X))$ (with respect to Lebesgue measure). The entropy numbers for this class satisfy $\epsilon_n(b(W^s(L_\infty(X))) = O(n^{-s/d})$. Thus, if we assume $f_\rho \in \Theta$ and take $\mathcal{H} = \Theta$, then (1.53) and (1.54) are valid with r replaced by s/d. We can improve this by taking the larger space $W^s(L_p(X))$, $p > d$, in place of $W^s(L_\infty(X))$. This class has the same asymptotic behavior of its entropy

numbers for its finite balls, and therefore whenever $f_\rho \in \Theta := b(W^s(L_p(X)),\ p > d$, then taking $\mathcal{H} = \Theta$, we have

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le Cm^{-\frac{s}{2s+2d}}, \quad m = 1, 2, \dots. \tag{1.55}$$

We stress that the spaces $W^s(L_p(X))$ are defined with respect to Lebesgue measure; they do not see the measure ρ_X.

1.6 The results of this paper

The purpose of the present paper is to make a systematic study of the rate of decay of learning algorithms as the number of samples increases and to understand what types of estimators will result in the best decay rates. In particular, we are interested in understanding what is the best rate of decay we can expect under a given prior $f_\rho \in \Theta$.

There are two sides to this story. The first is to establish lower bounds for the decay rate under a given prior. We let $\mathcal{M}(\Theta)$ be the class of all Borel measures ρ on Z such that $f_\rho \in \Theta$. Recall that we do not know ρ so that the best we can say about it is that it lies in $\mathcal{M}(\Theta)$. We enter into a competition over all estimators $I\!E_m : \mathbf{z} \to f_{\mathbf{z}}$ and define

$$e_m(\Theta) := \inf_{I\!E_m} \sup_{\rho \in \mathcal{M}(\Theta)} E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|_{L_2(X,\rho_X)}). \tag{1.56}$$

We note that in regression theory they usually study $E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|^2_{L_2(X,\rho_X)})$. From our probability estimates we can derive estimates for $E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|^q_{L_2(X,\rho_X)})$ for the whole range $1 \le q < \infty$. For the sake of simplicity we formulate our expectation results only in the case $q = 1$.

We give in §3 a method to obtain lower bounds for $e_m(\Theta)$ for a variety of different choices for the priors Θ. The main ingredients in this lower bound analysis are a different type of entropy (called *tight entropy*) and the use of concepts from information theory such as the Kullback-Leibler information and Fano inequalities. As an example, we recover the following result of Stone (see Theorem 3.2 in [19]): for $\Theta := b(W^s(L_p(X)))$,

$$e_m(b(W^s(L_p(X))) \ge c_s m^{-\frac{s}{2s+d}}, \quad m = 1, 2, \dots. \tag{1.57}$$

Notice that the best estimate we have obtained so far in (1.55) does not give this rate of decay.

We determine lower bounds for many other priors Θ. For example, we determine lower bounds for all the clasical Sobolev and Besov smoothness spaces. We phrase our analysis of lower bounds in such a way that it can be applied to non classical settings. It is our contention that the correct prior classes to analyze in learning should be smoothness (or approximation) classes that depend on ρ_X and we have formulated our analysis so as to possibly apply to such situations.

One of the points of emphasis of this paper is to formulate the learning problem in terms of probability estimates and not just expectation estimates. In this direction, we are following the lead of Cucker and Smale [10]. We shall now give a formal way to measure the performance of algorithms in probability which can be a useful benchmark.

Given our prior Θ and the associated class $\mathcal{M}(\Theta)$ of measures, we define for each $\eta > 0$ the *accuracy confidence function*

$$\mathbf{AC}_m(\Theta, \eta) := \inf_{I\!E_m} \sup_{\rho \in \mathcal{M}(\Theta)} \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| > \eta\}. \tag{1.58}$$

We shall prove lower bounds for **AC** of the following form

$$\mathbf{AC}_m(\Theta, \eta) \geq C \min(1/2, \sqrt{\bar{N}(\Theta, \eta)} e^{-cm\eta^2}). \tag{1.59}$$

Let η_m^* be the value of η where the two terms in the minimum occuring in (1.59) are equal. Then, this minimum is $1/2$ for $\eta \leq \eta_m^*$ and then the exponential term dominates. One can incorporate the term $\sqrt{\bar{N}(\Theta, \eta)}$ into the exponential and thereby obtain (see Figure 1.1)

$$\mathbf{AC}_m(\Theta, \eta) \geq C' \begin{cases} e^{-c'm\eta^2} & \eta \geq \eta_m^*(\Theta)/2 \\ 1, & \eta \leq \eta_m^*(\Theta)/2 \end{cases} \tag{1.60}$$

for appropriately chosen constants c', C'.

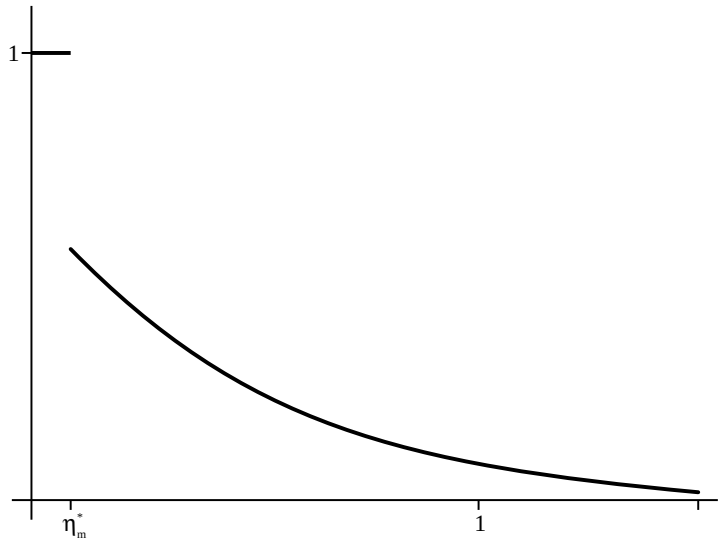

Figure 1.1: The typical graph of an **AC** function.

These lower bounds for **AC** are our vehicle for proving expectation lower bounds. We obtain the expectation lower bound $e_m(\Theta) \geq C\eta_m^*(\Theta)$.

The use of Kullback-Leibler information together with Fano inequalities is well known in statistics and goes back to Le Cam [27] and Ibragimov and Hasminskii [21] (see also e.g. [20]). What seems to separate our results from previous works is the generality in which this approach can be executed and the fact that our bounds (lower bounds and upper bounds) are obtained in terms of probability which go beyond bounds for the expected error.

The major portion of this paper is concerned with establishing upper bounds for $e_m(\Theta)$ and related probabilities and to understand what types of estimators will yield good upper

bounds. Typically we shall construct estimators that do not depend on η and show that they yield upper bounds for $\mathbf{AC}_m(\Theta,\eta)$ that have the same graphical behavior as in Figure 1.1:

$$\mathbf{AC}_m(\Theta,\eta) \le \begin{cases} Ce^{-cm\eta^2} & \eta \ge \eta_m(\Theta) \\ 1, & \eta \le \eta_m(\Theta). \end{cases} \tag{1.61}$$

By integrating such probabilistic upper bounds we derive the upper bound $e_m(\Theta) \le C\eta_m(\Theta)$. Notice that if $\eta_m(\Theta)$ and $\eta_m^*(\Theta)$ are comparable then we have a satisfactory description of $\mathbf{AC}_m(\Theta,\eta)$ save for the constants c, C.

It is possible to give estimators which provide upper bounds (both in terms of expectation and probability) that match the lower bounds for all of the Sobolev and Besov classes that are compactly embedded into $\mathcal{C}(X)$. The way to accomplish this is to use hypothesis classes $\mathcal{H}$ with smaller entropy. For example, choosing for each class a proper ϵ-net (depending on m) will do the job. This is shown in the follow up paper [25]. In particular, it implies the corresponding expectation estimates. Apparently, as was pointed out to us by Lucien Birgé, similar expectation estimates can also be derived from the results in [4]. Also, we should mention that for the Sobolev classes $W^k(L_\infty(X))$ and expectation estimates, this was also proved by Stone (see again [19]).

The ϵ net approach, while theoretically powerful, is not numerically implementable. We shall be interested in using other methods to construct estimators which may prove to be more numerically friendly. In particular, we want to see what we can expect from estimators based on other methods of approximation including widths and nonlinear approximation.

The estimation algorithms we construct in this paper will choose a hypothesis space $\mathcal{H}$ (which will generally depend on m) and take for $f_{\mathbf{z}}$ the empirical least squares estimator (1.26) to the data $\mathbf{z}$ from $\mathcal{H}$. There will be two types of choices for $\mathcal{H}$:

Prior dependent estimators: These will start with a prior class Θ and construct an estimator using the knowledge of Θ. Such an estimator, tailored to Θ will typically not perform well on other prior classes.

Prior independent estimators: These estimators will be built independent of any prior classes with the hope that they will perform well on a whole bunch of prior classes.

We shall say that an estimation algorithm $(I\!E_m)$ is *universally convergent* if $f_{\mathbf{z}}$ converges in expectation to f_ρ for each Borel measure ρ on X. Such algorithms are sometimes called consistent in statistics. We shall say that the algorithm is *optimal in expectation* for the prior class Θ if

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le C(\Theta)e_m(\Theta), \quad m = 1,2,\dots. \tag{1.62}$$

We say that the algorithm is optimal in probability if

$$\sup_{\rho\in\mathcal{M}(\Theta)} \rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| > \delta\} \le C_1\mathbf{AC}_m(\Theta, C_2\delta), \quad m = 1,2,\dots,\ \delta > 0, \tag{1.63}$$

with C_1, C_2 constants that may depend on Θ. We say that a learning algorithm is *universally optimal* (in expectation or probability) for a class $\mathcal{P}$ of priors if it is optimal for each $\Theta \in \mathcal{P}$. We shall often construct estimators which are not optimal because of the appearance of an additional logarithmic term $(\log m)^\nu$ for some $\nu > 0$ in the case of

expectation estimates. We shall call such estimators *semi-optimal*. This is in particular the case when we construct estimators that are effective for a wide class of priors (see e.g. §4.4): estimators that are effective for large classes of priors are called adaptive in the statistics literature.

The simplest example of the type of upper bounds we establish is given in Theorem 4.1 which uses Kolmogorov n-widths to build prior dependent estimators. If Θ is a compact set in $\mathcal{C}(X)$ whose Kolmogorov widths satisfy $d_n(\Theta, \mathcal{C}(X)) \le cn^{-r}$, then we choose $\mathcal{H}$ as $\mathcal{L}_n \cap b(\mathcal{C}(X))$, where $\mathcal{L}_n$ is a near optimal n dimensional subspace for Θ. We show in §4.1 that when $n := (\frac{m}{\ln m})^{\frac{1}{2r+1}}$ this will give an estimator $f_{\mathbf{z}} = f_{\mathbf{z},\mathcal{H}}$ such that

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le C(\frac{\ln m}{m})^{\frac{r}{2r+1}}, \tag{1.64}$$

with C a constant depending only on r. That is, these estimators are semi-optimal in their expectation bounds. A corresponding inequality in probability is also established. The estimate (1.64) applies to finite balls in Sobolev spaces, i.e. $\Theta = b(W^s(L_p))$ in which case $r = s/d$. It is shown that this again produces estimators which are semi-optimal provided $s > d/2$ and $p \ge 2$. The logarithm can be removed (for the above mentioned Sobolev spaces) by other methods (see [25] and in the case of expectation estimates Chapter 19 of [19]).

One advantage of using Kolmogorov widths is that with them we can construct a universal estimator. For example, we show in §4.4 that there is a single estimator that gives the inequalities (1.64) provided $a \le r < b$ with $a > 0$ an arbitrary but fixed constant. The constant b can be chosen arbitrarily in case of estimation in expectation but we only establish this for $r \le 1/2$ for estimates in probability. It remains an open problem whether this restriction on r can be removed in the case of probability estimates.

Another method for constructing universal estimators based on adaptive partitioning is given in [5]. The estimator there is semi-optimal for a range of Besov spaces with smoothness less than one (a restriction which comes about because the method uses piecewise constants for the construction of $f_{\mathbf{z}}$).

In §4.2, we show how to use nonlinear methods to construct estimators. These estimators can be considered as generalizations of thresholding operators based on wavelet decompositions. Recall that thresholding has proven to be very effective in a variety of settings in statistical estimation [17, 18]. For a range of Besov spaces, these esimators are proven to be semi-optimal.

In summary, as pertains to upper bounds, this paper puts forward a variety of techniques to obtain upper bounds and discusses their advantages and disadvantages. In some cases these estimators provide semi-optimal upper bounds. In some cases they can be modified (as reported on in subsequent papers) to obtain optimal upper bounds. We also highlight partial results on obtaining universally optimal estimators which we feel is an important open problem.

In (§5) we consider a variant of the learning problem in which we approximate a variant f_μ of f_ρ. Namely, we assume that $d\rho_X = \mu\, dx$ is absolutely continuous and approximate the function $f_\mu := \mu f_\rho$ from the given data $\mathbf{z}$. We motivate our interest in this function f_μ with the above banking problem. One advantage gained in estimating f_μ is that we can provide estimates in L_p without having to go through L_∞.

We consider the results of this paper to be theoretical but some of the methods put forward could potentially be turned into numerical methods. At this point, we do not address the numerical feasibility of our algorithms. Our main interest is to understand what is the best performance we can expect (in terms of accuracy-confidence or expected error decay with m) for the regression problem with various linear and nonlinear methods.

The use of entropy has a long history in statistical estimation. The use of entropy as proposed by Cucker and Smale [10] and also used here is similar in both flavor and execution to other uses in statistics (see for example the articles [4], the book of Sara van de Geer [38] or the book of Györfi, Kohler, Krzyżak, and Walk [19]). We have tried to explain the use of these concepts in a fairly accessible way, especially for researchers from the variious communities that relate to learning (statistics, functional analysis, probability and approximation) and moreover to show how other concepts of approximation such as Kolmogorov widths or nonlinear widths can be employed in learning. They have some advantages and disadvantages that we shall point out.

2 Priors described by smoothness or approximation properties

The purpose of this section is to introduce the types of prior sets Θ that we shall employ. Since we are interested in priors for which $e_m(\Theta)$ tends to zero as m tends to infinity, we must necessarily have Θ compact in $L_2(X, \rho_X)$ for each $\rho \in \mathcal{M}(\Theta)$. It is well known that compact subsets in L_p spaces (or $\mathcal{C}(X)$) have a uniform smoothness when measured in that space. Therefore, they are typically described by smoothness conditions. Another way to describe compact sets is through some type of uniform approximability of the elements of Θ. We shall use both of these approaches to describe prior sets. These two ways of describing priors are closely connected. Indeed, a main chapter in approximation theory is to characterize classes $\mathcal{A}$ of functions which have a prescribed approximation rate by showing that $\mathcal{A}$ is a certain smoothness space. Space will not allow us to describe this setting completely - in fact it is a subject of several books. However, we wish to present enough discussion for the reader to understand our viewpoint and to be able to understand the results we put forward in this paper. The reader may wish to skim over this section and return to it only as necessary to understand our results on learning theory.

2.1 Smoothness spaces

We begin by discussing smoothness spaces in $\mathcal{C}(X)$ or in $L_p(X)$ equipped with Lebesgue measure. This is a classical subject in mathematical analysis. The simplest and best known smoothness spaces are the Sobolev spaces $W^k(L_p(X))$, $1 \le p \le \infty$, $k = 1, 2, \ldots$. The space $W^k(L_p(X))$ is defined as the set of all functions $g \in L_p(X)$ whose distributional derivatives $D^\nu g$, $|\nu| = k$, are also in L_p. The semi-norm on this space is

$$|g|_{W^k(L_p(X))} := \sum_{|\nu|=k} \|D^\nu g\|_{L_p(X)}. \tag{2.1}$$

We obtain the norm $\|g\|_{W^k(L_p(X))}$ for this space (and all other smoothness spaces in $L_p(X)$) by adding $\|g\|_{L_p(X)}$ to the semi-norm.

The family of Sobolev spaces is insufficient for most problems in analysis because of two reasons. The first is that we would like to measure smoothness of order s when $s > 0$ is not an integer. The second is that in some cases, we want to measure smoothness in $L_p(X)$ with $p < 1$. There are several ways to define a wider family of spaces. We shall use the Besov spaces because they fit best with approximation and statistical estimation.

A Besov space $B^s_q(L_p(X))$ has three parameters. The parameter $0 < p \leq \infty$ plays the same role as in Sobolev spaces. It is the $L_p(X)$ space in which we measure smoothness. The parameter $s > 0$ gives the smoothness order and is the analogue of k for Sobolev spaces. The parameter $0 < q \leq \infty$ makes subtle distinctions in these spaces.

The usual definition of Besov spaces is made by either using moduli of smoothness or by using Fourier transforms and can be found in many texts (we also refer to the paper [15]). For example, for $0 < s < 1$, and $p = \infty$, the Besov space $B^s_\infty(L_\infty(X))$ is the same as the Lipschitz space Lip s whose semi-norm is defined by

$$|f|_{Lip\, s} := \sup_{x_1, x_2 \in X} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|^s}. \tag{2.2}$$

We shall not give the general definition of Besov spaces in terms of moduli of smoothness or Fourier transforms but rather give, later in this section, an **equivalent definition** in terms of wavelet decompositions (see §2.2) since this latter description is useful for understanding some of our estimation theorems using wavelet decompositions.

It is well known when a finite radius ball $b(W)$ of a Sobolev or Besov spaces W is compactly embedded in $L_p(X)$. This is connected to what are called Sobolev embedding theorems. To describe these results, it will be convenient to have a pictorial description of smoothness spaces. We shall use this pictorial description often in describing our results. We shall identify smoothness spaces with points in the upper right quadrant of $I\!R^2$. We write each such point as $(1/p, s)$ and identify this point with a smoothness space of smoothness order s in L_p; this space may be the Sobolev space $W^k(L_p(X))$ in the case $s = k$ is an integer or the Besov space $B^s_q(L_p(X))$ in the general case $s \geq 0$. The points $(1/p, 0)$ correspond to $L_p(X)$ when $p < \infty$ and to $\mathcal{C}(X)$ when $p = \infty$. The compact subsets of $L_p(X)$ are easy to describe using this picture. We fix the value of p and we consider the line segment whose coordinates $(1/\mu, s)$ satisfy $1/\mu = \frac{s}{d} + \frac{1}{p}$. This is the so-called Sobolev embedding line for $L_p(X)$. For any point $(1/\tau, s)$ to the left of this line, any finite ball in the corresponding smoothness space is compactly embedded in $L_p(X)$. Figure 2.1 depicts the situation for $p = \infty$, i.e. the spaces compactly embedded in $\mathcal{C}(X)$.

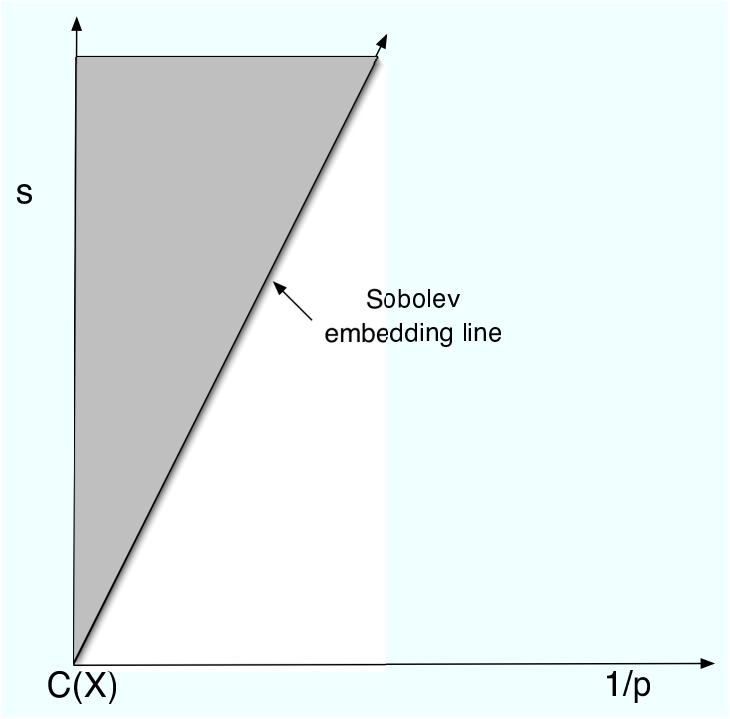

Figure 2.2: The shaded region depicts the smoothness spaces embedded in $\mathcal{C}(X)$ when $d = 2$. The Sobolev embedding line has equation $s = 2/p$ in this case.

Thus far, we have only described isotropic smoothness spaces, i.e. the smoothness is the same in each coordinate. There are also important anistropic spaces which measure smoothness differently in the coordinate directions. We describe one such family of smoothness spaces known as the Hölder-Nikol'skii classes $NH_p^{\mathbf{s}}$, defined for $\mathbf{s} = (s_1, \ldots, s_d)$ and $1 \leq p \leq \infty$. This class is the set of all functions $f \in L_p(X)$ such that for each $l_j = [s_j] + 1$, $j = 1, \ldots, d$, we have

$$\|f\|_p \leq 1, \qquad \|\Delta_t^{l_j,j} f\|_{L_p(X)} \leq |t|^{s_j}, \quad j = 1, \ldots, d, \tag{2.3}$$

where $\Delta_t^{l,j}$ is the l-th difference with step size t in the variable x_j. In the case $d = 1$, $NH_p^{\mathbf{s}}$ coincides with the standard Lipschitz ($0 < s < 1$) or Hölder ($s \geq 1$) classes. If $s_1 = \ldots = s_d = s$ these classes coincide with the Besov classes $B_\infty^s(L_p(X))$ (see §2.2).

2.2 Wavelet decompositions

In this section, we shall introduce wavelets and wavelet decompositions. These will be important in the construction of estimators later in this paper. Also, we shall use them to define the Besov spaces. There are several books which discuss wavelet decompositions and their characterization of Besov spaces (see e.g. Meyer [30] or the survey [16]). We also refer to the article of Daubechies [13] for the construction of wavelet bases of the type we want to use.

Let φ be a univariate scaling function which generates a univariate wavelet ψ which has compact support . For specificity, we take φ and ψ to be one of the Daubechies' pairs (see [13]) which generate orthogonal wavelets. We define $\psi^0 := \varphi$ and $\psi^1 := \psi$. Let E' be the set of vertices of the unit cube $[0,1]^d$ and let $E := E' \setminus \{(0,\dots,0)\}$ be the set of nonzero vertices. We also let $\mathcal{D}$ denote the set of dyadic cubes in $\mathbb{R}^d$ and $\mathcal{D}_j$ the set of dyadic cubes of side length 2^{-j}. Each $I \in \mathcal{D}_j$ is of the form

$$I = 2^{-j}[k_1, k_1+1] \times \cdots \times 2^{-j}[k_d, k_d+1], \quad k = (k_1,\dots,k_d) \in \mathbb{Z}^d. \tag{2.4}$$

For each $0 < p \le \infty$, the wavelet functions

$$\psi_I^e(x) := \psi_{I,p}^e(x) := 2^{jd/p}\psi^{e_1}(2^j x_1 - k_1)\cdots\psi^{e_d}(2^j x_d - k_d), \quad I \in \mathcal{D},\ e \in E, \tag{2.5}$$

(normalized in $L_p(\mathbb{R}^d)$) form an orthogonal system. Each locally integrable function f defined on $\mathbb{R}^d$ has a wavelet decomposition

$$f = \sum_{I\in\mathcal{D}}\sum_{e\in E} f_I^e \psi_I^e, \quad f_I^e := f_{I,p}^e := \langle f, \psi_{I,p'}^e\rangle, \quad 1/p + 1/p' = 1. \tag{2.6}$$

Here $f_I^e = f_{I,p}^e$ depends on the p-normalization that has been chosen but $f_I^e\psi_I^e$ is the same regardless of p. We shall usually be working with L_2 normalized wavelets. If this is not the case, we shall indicate the dependence on p. The series (2.6) converges absolutely to f in the $L_p(\mathbb{R}^d)$ norm in the case $f \in L_p(\mathbb{R}^d)$ and $1 < p < \infty$ and conditionally in the case $p = \infty$ with $L_\infty(\mathbb{R}^d)$ replaced by $C(\mathbb{R}^d)$.

For any cube J, we shall denote by $\ell(J)$ the side length of J. The wavelet functions ψ_I^e all have compact support. We take $\bar{I}$ as the smallest cube that contains the support of this wavelet. Then,

$$\ell(\bar{I}) \le A_0 \ell(I), \tag{2.7}$$

where A_0 depends only on the initial choice of the wavelet ψ.

In order to define Besov spaces in terms of wavelet coefficients, we let k be a positive integer such that the mother wavelet ψ is in $C^k(\mathbb{R}^d)$ and has k vanishing moments. If $0 < p, q \le \infty$ and $0 \le s < k$, then, for the p normalized basis $\{\psi_I^e\}$,

$$|f|_{B_q^s(L_p(\mathbb{R}^d))} := \begin{cases} \left(\sum_{j=-\infty}^{\infty} 2^{jsq}\left(\sum_{I\in\mathcal{D}_j}\sum_{e\in E}|f_I^e|^p\right)^{q/p}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{-\infty<j<\infty} 2^{js}\left(\sum_{I\in\mathcal{D}_j}\sum_{e\in E}|f_I^e|^p\right)^{1/p}, & q = \infty. \end{cases} \tag{2.8}$$

defines the (quasi-semi)-norm for the Besov space $B_q^s(L_p(\mathbb{R}^d))$. These spaces and norms are equivalent for different choices of k (provided $s < k$) and are equivalent to the classical definitions using moduli of smoothness as long as $1 \le p < \infty$ or $0 < p < 1$ and $s/d > 1/p - 1$. The quasi-norm in $B_q^s(L_p(\mathbb{R}^d))$ is defined by

$$\|f\|_{B_q^s(L_p(\mathbb{R}^d))} := \|f\|_{L_p(\mathbb{R}^d)} + |f|_{B_q^s(L_p(\mathbb{R}^d))}. \tag{2.9}$$

If s is not an integer, we define $W^s(L_p(X)) := B_p^s(L_p(X))$ which serves to extend the scale of Sobolev spaces to all s.

The wavelet decomposition (2.6) runs over all dyadic levels. There is an analogous decomposition that runs only over $j \geq j_0$ where j_0 is any fixed integer. Let $\mathcal{D}_+ := \cup_{j \geq j_0} \mathcal{D}_j$. Each locally integrable function on $\mathbb{R}^d$ has the wavelet decomposition

$$f = \sum_{j \geq j_0} \sum_{I \in \mathcal{D}_j} \sum_{e \in E_j} f_I^e \psi_I^e, \quad f_I^e := \langle f, \psi_{I,p'}^e \rangle. \tag{2.10}$$

where $E_{j_0} := E'$, and $E_j := E$, $j > j_0$. Here, for the case $j = j_0$, the wavelets ψ_I^e are replaced by the corresponding scaling functions φ_I^e.

We can also describe Besov norms using this decomposition:

$$\|f\|_{B_q^s(L_p(\mathbb{R}^d))} := \begin{cases} \left(\sum_{j=j_0}^{\infty} 2^{jsq} \left(\sum_{I \in \mathcal{D}_j} \sum_{e \in E} |f_I^e|^p \right)^{q/p} \right)^{1/q}, & 0 < q < \infty, \\ \sup_{j \geq j_0} 2^{js} \left(\sum_{I \in \mathcal{D}_j} \sum_{e \in E} |f_I^e|^p \right)^{1/p}, & q = \infty. \end{cases} \tag{2.11}$$

There are also wavelet decompositions for domains $\Omega \subset \mathbb{R}^d$. For our purposes, it will be sufficient to describe such a basis for $\Omega = [0,1]^d$. We start with a usual wavelet basis for $\mathbb{R}$ and construct a basis for $[0,1]$. The bases for $[0,1]$ will contain all of the usual $\mathbb{R}$ wavelet basis functions when these basis functions have supports strictly contained in $[0,1]$. The other wavelets in this basis are obtained by modifying the $\mathbb{R}$ wavelets whose supports overlap $[0,1]$ but are not contained completely in $[0,1]$. Notice that on any dyadic level there are only a finite number of wavelets that need to be modified. For details on this construction see [6]. To get the basis $[0,1]^d$ we take the tensor product of the $[0,1]$ basis. For the $[0,1]^d$ wavelet system one has the same characterization of Besov spaces as given above.

2.3 Approximation spaces

Another way to describe priors is by imposing decay conditions on rates of approximation. One situation that we have already encountered is to impose a condition on entropy numbers. For example, for $r > 0$, we can consider a class Θ such that

$$\epsilon_n(\Theta) \leq Cn^{-r}. \tag{2.12}$$

Such conditions are closely related to smoothness.

Let us first describe this for entropy conditions in $\mathcal{C}(X)$. For the Sobolev spaces $W^k(L_p(X))$ (with respect to Lebesgue measure), we have

$$\epsilon_n(b(W^k(L_p(X)))_{\mathcal{C}(X)} \leq Cn^{-k/d}, \quad n = 1, 2, \ldots, \tag{2.13}$$

provided $k > d/p$. Equivalently, this can be stated as

$$N(\delta, b(W^k(L_p(X)))_{\mathcal{C}(X)} \leq Ce^{c_0 \delta^{-\frac{d}{k}}}, \quad \delta > 0. \tag{2.14}$$

Similar results hold for any of the Lipschitz or Besov spaces which are compactly embedded into $\mathcal{C}(X)$. If $s > 0$ and $B_q^s(L_\tau(X))$ is a Besov space corresponding to a point to the left of the Sobolev embedding line for $\mathcal{C}(X)$ then

$$\epsilon_n(b(B_q^s(L_\tau(X)))_{\mathcal{C}(X)} \leq Cn^{-s/d}, \quad n = 1, 2, \ldots, \tag{2.15}$$

where the constant C depends on the distance of $(1/\tau, s)$ to the embedding line. We shall also use priors that utilize Kolmogorov widths in place of entropy numbers. These are formulated in §4.1.

2.4 Approximation using a family of linear spaces

We introduce in this subsection a typical way of describing compact sets using approximation. Let $\mathcal{B}$ be a Banach space and let $\mathcal{L}_n$, $n = 1, 2, \ldots$, be a sequence of linear subspaces of $\mathcal{B}$ with $\mathcal{L}_n$ of dimension $\leq n$. For simplicity we assume that $\mathcal{L}_n \subset \mathcal{L}_{n+1}$. Typical choices for $\mathcal{B}$ are the $L_p(X)$ spaces with respect to Lebesgue measure. Possible choices for $\mathcal{L}_n$ are space of algebraic or trigonometric polynomials. Note, that to maintain the condition on the dimension of $\mathcal{L}_n$, we would need to repeat these spaces of polynomials.

In this setting, we define for $f \in \mathcal{B}$

$$E_n(f) := E_n(f)_{\mathcal{B}} := \inf_{g \in \mathcal{L}_n} \|f - g\|_{\mathcal{B}} \tag{2.16}$$

which is the error in approximating f in the norm of $\mathcal{B}$ when using the elements of $\mathcal{L}_n$.

For any $r > 0$, we define the approximation class $\mathcal{A}^r := \mathcal{A}^r(\mathcal{B}, (\mathcal{L}_n))$ to be the set of all $f \in \mathcal{B}$ such that

$$|f|_{\mathcal{A}^r} := \sup_n n^r E_n(f). \tag{2.17}$$

The functions in $\mathcal{A}^r$ can be approximated to accuracy $|f|_{\mathcal{A}^r} n^{-r}$ when using the elements of $\mathcal{L}_n$.

There are slightly more sophisticated approximation classes $\mathcal{A}^r_q$ which make subtle distinctions in approximation order through the index $q \in (0, \infty]$. The seminorms for these classes are defined by

$$|f|_{A^r_q} := \|(n^r E_n(f))\|_{\ell_q *}, \tag{2.18}$$

where $\|(\alpha_n)\|^q_{\ell_q *} := \sum_{n=1}^{\infty} |\alpha_n|^q \frac{1}{n}$ when $q < \infty$ and is the usual ℓ_∞ norm when $q = \infty$.

A typical prior on the functions f_ρ is to assume that f_ρ is in a finite ball in $\mathcal{A}^r_q(\mathcal{B}, (\mathcal{L}_n))$ for a specific family of approximation spaces. The advantage of such priors over priors on smoothness spaces is they can be defined for $\mathcal{B} = L_2(\mu)$ for arbitrary μ.

A large and important chapter of approximation theory characterizes approximation spaces as smoothness spaces in the case approximation takes place in $L_p(X)$ with respect to Lebesgue measure. For example, consider the case of approximating 2π-periodic functions on $\mathbb{T}^d$ by trigonometric polynomials of degree $\leq n$ which is a linear space of dimension $(2n+1)^d$. Then, for any $1 \leq p \leq \infty$ (with the case $p = \infty$ corresponding to $\mathcal{C}$), we have

$$\mathcal{A}^r_q((\mathcal{L}_n), L_p(X))) = \odot B^{rd}_q(L_p), \quad r > 0,\ 0 < q \leq \infty, \tag{2.19}$$

where the $\odot B$ indicates we are dealing with periodic functions. A similar result holds if we replace trigonometric polynomials by spline functions of degree k on dyadic partitions with $k \geq s - 1$. The corresponding results for wavelet approximation will be discussed in the following subsection. The characterizations (2.19) provide a useful way of characterizing Besov spaces. It also shows that for many approximation methods the approximation classes are identical.

2.5 Approximation using orthogonal systems

We discuss in this subsection the important case where approximation comes from an orthonormal system (we could equally well consider Riesz bases). Let us suppose that $\Psi := \{\psi_j\}_{j=1}^{\infty}$is a complete orthonormal system for $L_2(X)$ with respect to Lebesgue measure. The classical settings here are the Fourier and orthonormal wavelet bases. There are two types of approximation that we want to single out corresponding to linear and nonlinear methods.

Any integrable function g has an expansion

$$g = \sum_{j=1}^{\infty} c_j(g)\psi_j; \quad c_j(g) := \int_X g\psi_j dx. \tag{2.20}$$

For a function g, we define

$$S_n(g) := \sum_{j=1}^{n} c_j(g)\psi_j. \tag{2.21}$$

This is the orthogonal projection of g onto the first n terms of the orthogonal basis Ψ. It is a linear method of approximation in that, for each n, we approximate from the linear space $\mathcal{L}_n := \text{span}\{\psi_1, \dots, \psi_n\}$ The error we incur in such an approximation is

$$E_n(g)_p := \|g - S_n(g)\|_{L_p(X)}. \tag{2.22}$$

As we have already noted in the previous section for the Fourier or wavelet bases, the approximation classes $\mathcal{A}_q^r(L_p)$ are identical to the Besov spaces $B_q^s(L_p(X))$, $s = r/d$ (see §2). In the case of the Fourier basis under lexicographic ordering, it is known that the projector S_n is bounded on $L_p(X)$, $1 < p < \infty$. Therefore, for $r > 0$ and $1 < p < \infty$,

$$g \in \dot{B}_{\infty}^r(L_p(X)) \quad \text{iff} \quad \|g - S_n(g)\|_{L_p(X)} \le Cn^{-r/d}, \quad n = 1, 2 \dots, \tag{2.23}$$

and the constant C is comparable with the norm of g in $\dot{B}_{\infty}^r(L_p(X))$.

In the case of a wavelet orthonormal system (with their natural ordering from coarse to fine and lexicographic at a given dyadic scale), we have the same result as in (2.23) except that the functions are no longer required to be periodic and the range of r is restricted to $r \le r_0$ where r_0 depends on the smoothness and number of vanishing moments of the mother wavelet.

There is a second way that we can approximate g from the orthogonal system $\{\psi_j\}$ which corresponds to nonlinear approximation. We define Σ_n to be the set of all functions S which can be written as a linear combination of at most n of the ψ_j:

$$S = \sum_{j \in \Lambda} c_j\psi_j, \quad \#(\Lambda) \le n. \tag{2.24}$$

In numerical considerations, we want to restrict the indices in Σ_n in order to make the search for good approximations reasonable. We define $\Sigma_{n,a}$ as the set of S in (2.24) with the added restriction $\Lambda \subset \{1, \dots, n^a\}$. For $0 < p \le \infty$,we define the error

$$\sigma_{n.a}(f)_p := \inf_{S \in \Sigma_{n,a}} \|f - S\|_{L_p(X)}. \tag{2.25}$$

Now, let us consider the special case of approximation in $L_2(X)$. A best approximation from $\Sigma_{n,a}$ to a given g is simply given by

$$G_{n,a}(g) := \sum_{j \in \Gamma_{n,a}} c_j(g)\psi_j, \tag{2.26}$$

where $\Gamma_{n,a}$ is the set of indices corresponding to the n largest (in absolute value) coefficients $|c_j(g)|$ with $j \le n^a$ Here we do not have uniqueness because of possible ties in the size of the coefficients; these ties can be treated in any way to construct a $\Gamma_{n,a}$. Thus,

$$\sigma_{n,a}(g)_2^2 = \sum_{j \notin \Gamma_{n,a}} |c_j(g)|^2. \tag{2.27}$$

Another way to describe the process of creating best approximations from $\Sigma_{n,a}$ is by thresholding. If $\lambda > 0$, we denote by $\Gamma(g, \lambda, a)$ the set of those indices $j \le n^a$ such that $|c_j(g)| \ge \lambda$. Then,

$$T_{\lambda,a}(g) := \sum_{j \in \Gamma(\lambda,a,g)} c_j(g)\psi_j \tag{2.28}$$

is a best approximation from $\Sigma_{n,a}$ to g in $L_2(X)$ where $n := \#(\Gamma(\lambda, a, g))$.

It is also very simple, in the $L_2(X)$ approximation case, to describe the approximation classes. For example, a function $g \in \mathcal{A}^r_\infty(L_2(X))$, i.e. $\sigma_{n,a}(g)_2 \le C_0 n^{-r}$, $n = 1, 2, \ldots$, if and only if the following hold:

$$\#(\Gamma(\lambda, a, g)) \le C_1 \lambda^{-\tau}, \quad \lambda > 0, \ \frac{1}{\tau} = r + \frac{1}{2} \tag{2.29}$$

and

$$E_{n^a}(g)_2 \le C_1 n^{-r}, \quad n = 1, 2, \ldots \tag{2.30}$$

and the constants C_1 and C_0 are comparable.

A case of special interest to us will be when Ψ is a wavelet basis (see §2). In this case, the characterizations (2.29), (2.30) are related to Besov spaces. For example, whenever $g \in B^{rd}_\tau(L_\tau(X))$, the condition (2.29) is satisfied. As we have already noted, the condition (2.30) is characterized by $g \in B^{rd/a}_\infty(L_2(X))$. Because of the Sobolev embedding theorem, both conditions will be satisfied if $g \in B^{rd}_\mu(L_\mu(X))$ provided

$$r + \frac{1}{2} - \frac{1}{\mu} \ge \frac{r}{a}. \tag{2.31}$$

In other words, if the mother wavelet for Ψ is in C^k and has k vanishing moments, then we have

Remark 2.1 *If $a > 0$, then conditions (2.29) and (2.30) are satisfied for all $f \in B^{rd}_\mu(L_\mu(X))$ provided $rd < k$ and (2.31) holds.*

2.6 Universal methods of approximation

In evaluating a particular approximation process, one can look at the classes of functions for which the approximation process gives optimal or near optimal performance. For example, if we use a sequence $(\mathcal{L}_n)$ of linear spaces of dimension n, we say this sequence is *near optimal* for approximating the elements of the compact set K in the norm of the Banach space $\mathcal{B}$ if

$$\operatorname{dist}(K, \mathcal{L}_n)_{\mathcal{B}} \le C d_n(K, \mathcal{B}) \tag{2.32}$$

where d_n is the Kolmogorov width of K. The same notion can be given for nonlinear methods of approximation except now we would compare performance against nonlinear widths.

Some approximation systems are near optimal for a large collection of compact sets K. We say that a sequence $(\mathcal{L}_n)$ of linear spaces of dimension n are universally near optimal for the collection $\mathcal{K}$ of compact sets K if (2.32) holds for each $K \in \mathcal{K}$ with a universal constant $C > 0$. That is, the one sequence of linear space $(\mathcal{L}_n)$ is simultaneously near optimal for all these compact sets. There is the analogous concept of universally near optimal with respect to nonlinear methods. In this case, one replaces in (2.32) the linear space $\mathcal{L}_n$ by nonlinear spaces depending on n parameters and replaces the Kolmogorov width by the corresponding nonlinear width. In the learning problem we shall introduce a similar universal concept for learning algorithms. Therefore we want to briefly describe what is known about universality in the approximation setting for the purposes of comparison with our later results.

Let us begin the discussion by considering a wavelet system of compactly supported wavelets from $C^k(X)$ which have vanishing moments up to k. The Besov space $B_q^s(L_\tau)$ is compactly embedded in L_p if and only if $s > (d/\tau - d/p)_+$. For any fixed $\delta > 0$, let $\mathcal{K}$ be the set consisting of all unit balls $u(B_q^s(L_p))$ with $s - d/\tau + d/p \ge \delta$ and $0 < s < k$. Then, nonlinear wavelet approximation based on thresholding is near optimal for $L_p(X)$ approximation (Lebesgue measure) for all of the sets $K \in \mathcal{K}$.

The standard wavelet system is suitable only to approximate isotropic classes. It is a more subtle problem to find systems that are universal for both isotropic and anisotropic classes. We shall discuss this topic in the case of multivariate periodic functions.

We have introduced earlier in §2.1 the collection of anisotropic Hölder-Nikol'skii classes $NH_q^{\mathbf{s}}$. It is known (see for instance [36]) that the Kolmogorov n-widths of these classes behave asymptotically as follows:

$$d_n(NH_q^{\mathbf{s}}, L_q) \asymp n^{-g(\mathbf{s})}, \quad 1 \le q \le \infty, \tag{2.33}$$

where

$$g(\mathbf{s}) := (\sum_{j=1}^{d} s_j^{-1})^{-1}.$$

In the case of periodic functions, we can find for each $\mathbf{s}$ a near optimal subspace $\mathcal{L}_n$ of dimension n for $NH_q^{\mathbf{s}}$ in L_q, i.e. it satisfies (2.32). The space $\mathcal{L}_n$ can be taken for example as the set of all trigonometric polynomials with frequencies k satisfying the inequalities

$$|k_j| \le 2^{g(\mathbf{s})l/s_j}, \quad j = 1, \dots, d, \tag{2.34}$$

where l is the largest integer such that the number of vectors $\mathbf{k}$ satisfying the above inequalities is $\le n$.

Notice that the subspaces $\mathcal{L}_n$ described by (2.34) are different for different $\mathbf{s}$ and therefore do not satisfy our quest for a universally near optimal approximating method. For given $\mathbf{a}, \mathbf{b}$ with $0 < a_j < b_j$, $j = 1, \dots, d$, and a given p, we consider the class

$$\mathcal{K}_{q,p}([\mathbf{a}, \mathbf{b}]) := \{u(NH_q^{\mathbf{s}}) : \ a_j \le s_j \le b_j, \ j = 1, 2, \dots, d, \ g(\mathbf{a}) > (1/q - 1/p)_+\}. \quad (2.35)$$

Each of the sets in $\mathcal{K}_{p,p}$ is compact in $L_p(X)$. It can be shown that there does not exist a sequence $(\mathcal{L}_n)$ of linear space $\mathcal{L}_n$ of dimension n which is universally optimal for this collection of compact sets. In fact, it is proved in [36] that for a sequence of linear spaces $(\mathcal{L}_n)$ to satisfy (2.32) for $\mathcal{K}_{p,p}$ then one must necessarily have $\dim(\mathcal{L}_n) \ge c(\log n)^{d-1} n$. Moreover, this result is optimal in the sense that we can create a sequence of spaces with this dimension that satisfy (2.32).

If we turn to nonlinear methods then we can achieve universality for the class (2.35). We describe one such result. We consider the library $\mathcal{O}$ consisting of all orthonormal bases O on X. For each n and O, we consider the error $\sigma_n(f, O)_p$ of n term approximation when using the orthogonal basis O (see the definition (2.25) with $a = \infty$). Given a set K, we define

$$\sigma_n(K, O)_p := \sup_{f \in K} \sigma_n(f, O)_p. \quad (2.36)$$

and

$$\sigma_n(K, \mathcal{O})_p := \sup_{f \in K} \inf_{O \in \mathcal{O}} \sigma_n(f, O)_p. \quad (2.37)$$

We say that the basis O is near-optimal for the class K if

$$\sigma_n(K, O)_p \le C \sigma_n(K, \mathcal{O})_p, \quad n = 1, 2, \dots. \quad (2.38)$$

In analogy to the linear setting, we say that O is universally near-optimal for a collection $\mathcal{K}$ of compact sets K if (2.38) holds for all $K \in \mathcal{K}$ with an absolute constant. It is shown in [37] that there exists an orthogonal basis which is universally near-optimal for the collection $\mathcal{K}_{q,p}$ defined in (2.35) for $1 < q < \infty$, $2 \le p < \infty$. Also, for each $K = u(NH_q^{\mathbf{s}})$, $\sigma_n(K, \mathcal{O})_p \approx n^{-g(\mathbf{s})}$, $1 < q < \infty$, $2 \le p < \infty$.

3 Lower bounds

In this section, we shall establish lower bounds for the accuracy that can be attained in estimating the regression function f_ρ by any learning algorithm. We will establish our lower bounds in the case $X = [0,1]^d$, $Y = [-1, 1]$ and $Z = X \times Y$. In going further in this paper, these lower bounds will serve as a guide for us in terms of how we would like specific algorithms to perform.

We let Θ be a given set of functions defined on X which corresponds to the prior we assume for f_ρ. We define, as in the introduction, the class $\mathcal{M}(\Theta)$ of all Borel measures ρ on Z for which $f_\rho \in \Theta$ and define $e_m(\Theta)$ by (1.56). We shall even be able to prove lower bounds with weaker assumptions on the learning algorithms $I\!E_m$. Namely, in addition

to allowing the learning algorithm to know Θ, we shall also allow the algorithm to know the marginal ρ_X. To formulate this, we let μ be any Borel measure defined on X and let $\mathcal{M}(\Theta, \mu)$ denote the set of all $\rho \in \mathcal{M}(\Theta)$ such that $\rho_X = \mu$ and consider

$$e_m(\Theta, \mu) := \inf_{I\!E_m} \sup_{\rho \in \mathcal{M}(\Theta,\mu)} E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|_{L_2(X,\mu)}). \tag{3.1}$$

We shall give lower bounds for e_m and related probabilities. To prove these lower bounds we introduce a different type of entropy.

3.1 Tight entropy

We shall establish lower bounds for e_m in terms of a certain variant of the Kolmogorov entropy of Θ which we shall call *tight entropy*. This type of entropy has been used to prove lower bounds in approximation theory. Also, a similar type of entropy was used by Yang and Barron [41] in statistical estimation. The entropy measure that we shall use is in general different from the Kolmogorov entropy, but, as we shall show later, for classical smoothness sets Θ, it is equivalent to the Kolmogorov entropy and therefore our lower bounds will apply in these classical settings.

We assume that $\Theta \subset L_2(X, \mu)$. Let $0 < c_0 \le c_1 < \infty$, be two fixed real numbers. We define the *tight packing numbers*

$$\bar{N}(\Theta, \delta, c_0, c_1) := sup\{N : \exists\, f_0, f_1, ..., f_N \in \Theta, \text{ with } c_0\delta \le \|f_i - f_j\|_{L_2(X,\mu)} \le c_1\delta,\ \forall i \ne j\}. \tag{3.2}$$

We will use the abbreviated notation $\bar{N}(\delta) := \bar{N}(\Theta, \delta, c_0, c_1)$, when there is no ambiguity on the choice of the other parameters. Obviously, if Θ is a subset of a normed space, then for all $R > 0$, $\bar{N}(R\Theta, \delta, c_0, c_1) = \bar{N}(\Theta, \frac{\delta}{R}, c_0, c_1)$.

3.2 The main result

Let us fix any set Θ and any Borel measure μ defined on X. We set $\mathcal{M} := \mathcal{M}(\Theta, \mu)$ as defined above. We also take $c_0 < c_1$ in an arbitrary way but then fix these constants. For any fixed $\delta > 0$, we let $\{f_i\}_{i=0}^{\bar{N}}$, with $\bar{N} := \bar{N}(\delta)$, be a net of functions satisfying (3.2). To each f_i, we shall associate the measure

$$d\rho_i(x, y) := (a_i(x)d\delta_1(y) + b_i(x)d\delta_{-1}(y))d\mu(x), \tag{3.3}$$

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

We have the following theorem.

Theorem 3.1 *Let $0 < c_0 < c_1$ be fixed constants. Suppose that Θ is a subset of $L_2(\mu)$ with packing numbers $\bar{N} := \bar{N}(\delta) := \bar{N}(\Theta, \delta, c_0, c_1)$. In addition suppose that for $\delta > 0$, the net of functions $\{f_i\}_{i=0}^{\bar{N}}$ in (3.2) satisfies $\|f_i\|_{\mathcal{C}(X)} \le 1/4$, $i = 0, 1, \ldots, \bar{N}$. Then for any estimator $f_{\mathbf{z}}$ we have for $c_2 := e^{-3/e}$ and some $i \in \{0, 1, \ldots, \bar{N}\}$*

$$\rho_i^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_i\|_{L_2(X,\mu)} \ge c_0\delta/2\} \ge \min(1/2, c_2\sqrt{\bar{N}(\delta)}e^{-2c_1^2 m\delta^2}), \quad \forall \delta > 0,\ m = 1, 2, \ldots, \tag{3.4}$$

and for some $\rho \in \mathcal{M}(\Theta, \mu)$, we have

$$E_{\rho^m}(\|f_{\mathbf{z}} - f_\rho\|_{L_2(X,\rho_X)}) \geq c_0\delta^*/4, \tag{3.5}$$

whenever $\ln \bar{N}(\delta^*) \geq 4c_1^2 m(\delta^*)^2$.

The remainder of this subsection will be devoted to the proof of this theorem.

The first thing we wish to observe is that the measures ρ_i are close to one another. To formulate this, we use the Kullback-Leibler information. Given two probability measures dP and dQ defined on the same measure space and such that dP is absolutely continuous with respect to dQ, we write $dP = gdQ$ and define

$$\mathcal{K}(P, Q) := \int \ln g dP = \int g \ln g dQ. \tag{3.6}$$

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

It is obvious that

$$\mathcal{K}(P^m, Q^m) = m\mathcal{K}(P, Q). \tag{3.7}$$

Lemma 3.2 *For any Borel measure μ and the measures ρ_i defined by (3.3), we have*

$$\mathcal{K}(\rho_i, \rho_j) \leq \frac{16}{15}\|f_i - f_j\|^2_{L_2(X,\mu)}, \quad i, j = 0, \ldots, \bar{N}. \tag{3.8}$$

Proof: We fix i and j. We have $d\rho_i(x, y) = g(x, y)d\rho_j(x, y)$, where

$$g(x, y) = \frac{1 + (\operatorname{sign} y)f_i(x)}{1 + (\operatorname{sign} y)f_j(x)} = 1 + \frac{(\operatorname{sign} y)(f_i(x) - f_j(x))}{1 + (\operatorname{sign} y)f_j(x)}. \tag{3.9}$$

Thus,

$$2K(\rho_i, \rho_j) = \int_X F_{i,j}(x)d\mu(x) \tag{3.10}$$

where

$$F_{i,j}(x) := (1 + f_i(x)) \ln(1 + \frac{f_i(x) - f_j(x)}{1 + f_j(x)}) + (1 - f_i(x)) \ln(1 - \frac{f_i(x) - f_j(x)}{1 - f_j(x)}). \tag{3.11}$$

Using the inequality $\ln(1 + u) \leq u$, we obtain

$$\begin{aligned} F_{i,j}(x) &\leq (f_i(x) - f_j(x)) \left\{ \frac{1 + f_i(x)}{1 + f_j(x)} - \frac{1 - f_i(x)}{1 - f_j(x)} \right\} \\ &= \frac{2|f_i(x) - f_j(x)|^2}{1 - f_j(x)^2} \leq (32/15)|f_i(x) - f_j(x)|^2. \end{aligned}$$

Putting this in (3.10), we deduce (3.8). □

To prove the lower bound stated in Theorem 3.1, we shall use the following version of Fano inequalities which is a slight modification of that given by Birgé [41].

Lemma 3.3 *Let $\mathcal{A}$ be a sigma algebra on the space Ω. Let $A_i \in \mathcal{A}, \quad i \in \{0, 1, \ldots, n\}$ such that $\forall i \neq j, \quad A_i \cap A_j = \emptyset$. Let $P_i, \quad i \in \{0, 1 \ldots, n\}$ be $n+1$ probability measures on $(\Omega, \mathcal{A})$. If*

$$p := \sup_{i=0}^{n} P_i(\Omega \setminus A_i),$$

then either $p > \frac{n}{n+1}$ or

$$\inf_{j \in \{0,1,\ldots,n\}} \frac{1}{n} \sum_{i \neq j} \mathcal{K}(P_i, P_j) \geq \Psi_n(p), \tag{3.12}$$

where

$$\Psi_n(p) := (1-p) \ln \left(\frac{1-p}{p}\right)\left(\frac{n-p}{p}\right) - p \ln \left(\frac{n-p}{np}\right) = \ln n + (1-p) \ln \left(\frac{1-p}{p}\right) - p \ln \left(\frac{n-p}{p}\right). \tag{3.13}$$

Proof The proof of this lemma follows the same arguments as Birgé and therefore we shall only sketch the main steps. We begin with the following duality statement which holds for probability measures P and Q:

$$\mathcal{K}(P, Q) = \sup\{\int f dP, \quad \int \exp f dQ = 1\}. \tag{3.14}$$

This result goes back at least to the Sanov theorem (see a.e. Dembo-Zeitouni [12]). Taking $f = \lambda \chi_A$ in (3.14), we find that for all $A \in \mathcal{A}$ and $\lambda \in I\!R$, we have

$$\mathcal{K}(P, Q) \geq \lambda P(A) - \log[(\exp \lambda - 1) Q(A) + 1] = \lambda P(A) - \phi_{Q(A)}(\lambda), \tag{3.15}$$

where for $0 < q < 1, \quad \lambda \in I\!R$

$$\phi_q(\lambda) := \log[(\exp \lambda - 1) q + 1] = \log[q \exp \lambda + 1 - q]$$

Note that $\phi_q(\lambda)$ is convex in λ, while it is concave and nondecreasing in q if $\lambda \geq 0$.

If we apply (3.15) to P_i and P_0 for each $i = 1, \ldots, n$ and then sum we obtain

$$\frac{1}{n} \sum_{i=1}^{n} K(P_i, P_0) \geq \lambda \frac{1}{n} \sum_{i=1}^{n} P_i(A_i) - \frac{1}{n} \sum_{i=1}^{n} \phi_{P_0(A_i)}(\lambda). \tag{3.16}$$

Obviously, if $\lambda \geq 0$, then

$$\lambda \frac{1}{n} \sum_{i=1}^{n} P_i(A_i) \geq \lambda \inf_{i=0}^{n} P_i(A_i) = \lambda(1-p).$$

Using convexity and monotonicity, we have for $\lambda \in I\!R$

$$-\frac{1}{n} \sum_{i=1}^{n} \phi_{P_0(A_i)}(\lambda) \geq -\phi_{\frac{1}{n}\sum_{i=1}^{n} P_0(A_i)}(\lambda) = -\phi_{\frac{1}{n} P_0(\cup_{i=1}^{n} A_i)}(\lambda).$$

Using again the fact that $q \mapsto \phi_q(\lambda)$ is non decreasing, together with $P_0(\cup_{i=1}^n A_i) \leq (1 - P_0(A_0)) = P_0(A_0^c) \leq p$ gives that for $\lambda \geq 0$,

$$-\phi_{\frac{1}{n}P_0(\cup_{i=1}^n A_i)}(\lambda) \geq -\phi_{\frac{p}{n}}(\lambda).$$

Therefore, $\forall \lambda \geq 0$,

$$\frac{1}{n}\sum_{i=1}^n K(P_i, P_0) \geq \lambda(1-p) - \phi_{\frac{p}{n}}(\lambda)$$

To complete the proof, we define

$$\sup_{\lambda \geq 0}(\lambda t - \phi_q(\lambda) =: \phi_q^*(t).$$

One easily checks that

$$\phi_q^*(t) = \begin{cases} 0 & \text{if } \ t < q \\ t\log(\frac{t}{q}) + (1-t)\log(\frac{1-t}{1-q}) & \text{if } \ q \leq t \leq 1 \\ \infty & \text{if } \ t > 1. \end{cases}$$

We now take $q = p/n$ and $t = 1 - p$ and use the above in (3.16), we obtain

$$\frac{1}{n}\sum_{i=1}^n \mathcal{K}(P_i, P_0) \geq \phi_{p/n}^*(1-p). \tag{3.17}$$

We can replace P_0 by P_j for any $j \in \{0, 1, \ldots, n\}$ in the above argument. Using this we easily derive (3.13) which completes the proof of the lemma . □

Proof of Theorem 3.1 We define $A_i := \{\mathbf{z} : \|f_{\mathbf{z}} - f_i\|_{L_2(\mu)} < c_0\delta/2\}$, $i = 0, \ldots, \bar{N} = \bar{N}(\Theta, \delta)$ with c_0 the constant in (3.2). Then, the sets A_i are disjoint because of (3.2). We apply Lemma 3.3 with our measures ρ_i^m and find that either $p \geq 1/2$ or

$$2c_1^2 m\delta^2 \geq \Psi_{\bar{N}}(p) \geq -\ln p + (1-p)\ln \bar{N} + (1-p)\ln(1-p) + 2p\ln p \geq -\ln p + (1/2)\ln \bar{N} - 3/e, \tag{3.18}$$

where we have used that $x \ln x$ has the minimum value $-1/e$ on $[0, 1]$. From (3.18), we derive (3.4). Now given δ^* such that $\sqrt{\bar{N}(\delta^*)} \geq e^{2c_1 m(\delta^*)^2}$, we have from (3.4) that for this δ^* there is an i such that with $\rho = \rho_i$, we have

$$\rho^m(\|f_{\mathbf{z}} - f_\rho\|_{L_2(\rho_X)} > c_0\delta^*/2) \geq 1/2. \tag{3.19}$$

It follows that for any $\delta \leq \delta^*$, (3.19) also holds. Integrating with respect to δ we obtain (3.5). This completes the proof of the theorem. □

3.2.1 Lower bounds for Besov classes

In this subsection, we shall show how to employ Theorem 3.1 to obtain lower bounds for the learning problem with priors given as balls in Besov spaces (with these spaces defined

relative to Lebesgue measure). We first show how to obtain lower bounds for the prior $\Theta = b(B^s_q(L_\infty(X))$, $s > 0$, $0 < q \le \infty$. We shall take $X = [0,1]^d$ and $d\mu$ to be Lebesgue measure. From this, one can deduce the same lower bounds for any minimally smooth domains X with again $d\mu$ Lebesgue measure.

To construct an appropriate net for Θ we shall use tensor product B-splines on dyadic partitions. We fix a $\delta > 0$ and choose j as the smallest integer such that $2^{-js} \le \delta$. For any $j = 1, 2, \dots$ and for $k := \lceil s \rceil$, there are $\ge 2^{jd}$ tensor product B-splines of degree k at the dyadic level j. They each have support on a cube with side length $2^{-j}k$. We can choose $J \ge c2^{jd}$ of these B-splines with disjoint supports. We label these as $\{\phi_i\}_{i=1}^J$ and normalize them in $L_2(X)$. Then , $\|\phi_i\|_{L_\infty} \le c\sqrt{J}$.

We construct a net of functions f_i which satisfy (3.2). As was shown in [23], we can choose at least $e^{J/8}$ subsets $\Lambda_i \subset \{1, \dots, J\}$ such that for each i, j we have $\#((\Lambda_i \setminus \Lambda_j) \cup (\Lambda_j \setminus \Lambda_i)) \ge J/4$. For each such Λ_i, we define

$$f_i := \frac{\delta}{\sqrt{J}} \sum_{j \in \Lambda_i} \phi_j. \tag{3.20}$$

This net $\{f_i\}$ of functions satisfy

$$\delta/2 \le \|f_i - f_j\|_{L_2(\mu)} \le \delta. \tag{3.21}$$

Also,

$$\|f_i\|_{\mathcal{C}(X)} \le c\delta \tag{3.22}$$

where we used our remark on the supports of the ϕ_i. The inequality (3.22)means that our condition $\|f_i\|_{L_\infty(X)}$ of Theorem 3.1 will be satisfied provided $\delta < \delta_0$ for a fixed $\delta_0 > 0$.

We next want to show that each of the functions f_i is in Θ provided we take the radius of this ball sufficiently large (depending only on d). For this, we consider the approximation of a given function $f \in \mathcal{C}(X)$ by linear combinations of all tensor product B-splines from dyadic level n. If we denote by $E'_n(f)$ the error of approximation in $\mathcal{C}(X)$ to f by this space of splines , then we have

$$E'_n(f_i) \le c \begin{cases} \delta, & n \le j \\ 0, & n > j. \end{cases} \tag{3.23}$$

This means that

$$\sum_{n=1}^{\infty} [2^{ns} E'_n(f_i)]^q \le \delta^q \sum_{n=1}^{j} 2^{nsq} \le C^q \delta^q 2^{jsq} \le C^q, \tag{3.24}$$

where C depends only on q and d. The convergence of the sum in (3.24) is a characterization of the Besov space $B^s_q(L_\infty(X))$ by linear approximation as noted in §2.3.

We have just proven that $\bar{N}(\delta, B^s_q(L_\infty)) \ge e^{J/8}$ provided $\delta \le J^{-s/d}$. Equivalently, we have proved that $\bar{N}(\delta, \Theta) \ge c_3 e^{\delta^{-\frac{d}{s}}}$ for each $0 < \delta < \delta_0$. Let us now apply Theorem 3.1. Estimate (3.5) gives that

$$e_m(\Theta) \ge e_m(\Theta, d\mu) \ge c_0 \delta^*/4 \tag{3.25}$$

for any δ^* which satisfies $\ln \bar{N}(\delta^*) \geq 4c_1 m(\delta^*)^2 + 1$, i.e. provided $(\delta^*)^{-d/s} \geq cm(\delta^*)^2$. From this, we obtain

$$e_m(\Theta) \geq e_m(\Theta, d\mu) \geq cm^{-\frac{s}{2s+d}}, \quad m = 1, 2, \ldots. \tag{3.26}$$

A similar analysis shows that (3.4) gives that for any estimator $f_{\mathbf{z}}$,

$$\sup_{\rho \in \mathcal{M}(\Theta)} \rho^m\{\mathbf{z} : \|f_{\mathbf{z}} - f_\rho\|_{L_2(X,\rho_X)} \geq c\delta\} \geq \begin{cases} 1/2, & \delta \leq 2\delta^*, \\ Ce^{-c\delta^2 m}, & \delta > 2\delta^*, \end{cases} \tag{3.27}$$

where $\delta^* = cm^{-\frac{s}{2s+d}}$ is the turning value as described above. These are the lower bounds we want for the Besov space $B^s_q(L_\infty(X))$. Because each Besov space $B^s_q(L_p(X))$ contains the corresponding $B^s_q(L_\infty(X))$, we obtain the same lower bounds for these spaces.

4 Estimates for f_ρ

In this section, we shall introduce several methods for constructing estimators $f_{\mathbf{z}}$. Typically, we assume that $f_\rho \in \Theta$ where $\Theta = b(W)$ is a ball in a space W which is assumed to have a certain approximation property. We then use this approximation property to choose a set $\mathcal{H}$ and define the estimator $f_{\mathbf{z}} \in \mathcal{H}$ as the least squares fit to the data $\mathbf{z}$ from $\mathcal{H}$. We then prove an estimate for the rate that $f_{\mathbf{z}}$ approximates f_ρ (in $L_2(X, \rho_X)$). These estimates will typically give (save for a possible logarithmic term) the optimal rate for this class.

4.1 Estimates for classes based on Kolmogorov widths.

In this subsection, we shall assume that $\Theta \subset b_{R_0}(\mathcal{C}(X))$ for some R_0 and that its Kolmogorov widths (1.14) satisfy [6]

$$d_n(\Theta, \mathcal{C}(X)) \leq Cn^{-r}, \quad n = 1, 2, \ldots. \tag{4.1}$$

This means that for each n, there is a linear subspace $\mathcal{L}_n$ of $\mathcal{C}(X)$ of dimension n such that

$$\operatorname{dist}(\Theta, \mathcal{L}_n)_{\mathcal{C}(X)} \leq C_1 n^{-r}, \quad n = 1, 2 \ldots. \tag{4.2}$$

There is an inequality of Carl [7] that compares entropy to widths. It says that whenever (4.1) holds then

$$\epsilon_n(\Theta, \mathcal{C}(X)) \leq C_2 n^{-r} \quad n = 1, 2, \ldots. \tag{4.3}$$

Therefore, the prior $f_\rho \in \Theta$ is typically stronger than the corresponding assumption (1.46).

The following theorem shows that under the assumption (4.1), we can derive a better estimate than that given in Corollary 1.1

[6] We shall use the following convention about constants. Those constants whose value may be important later will be denoted with subscripts. Constants with no subscript such as c, C can vary with each occurrence even in the same line.

Theorem 4.1 *Let $f_\rho \in \Theta$ where $\Theta \subset b_{R_0}(\mathcal{C}(X))$ and Θ satisfies (4.2). Given $m \geq 2$, we take $n := (\frac{m}{\ln m})^{\frac{1}{2r+1}}$ and define $\mathcal{H} := \mathcal{H}_m := b_R(\mathcal{C}(X)) \cap \mathcal{L}_n$ where $R := M + C_1$. Then, the least squares estimator $f_{\mathbf{z}}$ for this choice of $\mathcal{H}$ satisfies*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \geq \eta\} \leq C \begin{cases} e^{-cm\eta^2} & \eta \geq \eta_m \\ 1, & \eta \leq \eta_m, \end{cases} \tag{4.4}$$

where $\eta_m := C(\ln m/m)^{\frac{r}{1+2r}}$ and the constants c, C depend only on C_1 and M. In particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \leq C(\frac{\ln m}{m})^{\frac{r}{2r+1}} \tag{4.5}$$

where C is also an absolute constant.

Proof: By our assumption, there is a $\phi_n \in \mathcal{L}_n$ such that $\|f_\rho - \phi_n\|_{L_\infty(X)} \leq C_1 n^{-r}$. Since $\|f_\rho\|_{L_\infty(X)} \leq M$, we have $\|\phi_n\|_{L_\infty(X)} \leq M + C_1 n^{-r}$. This gives that $\phi_n \in \mathcal{H}$ for our choice of $R = C_1 + M$. Therefore, with this choice of R, and $\mathcal{H} := b_R(\mathcal{C}(X)) \cap \mathcal{L}_n$, we have the estimate

$$\operatorname{dist}(f_\rho, \mathcal{H})_{\mathcal{C}(X)} \leq C_1 n^{-r}. \tag{4.6}$$

It follows that

$$\operatorname{dist}(f_\rho, \mathcal{H})_{L_2(X,\rho_X)} \leq C_1 n^{-r}. \tag{4.7}$$

For any $\eta > 0$, we have that the covering numbers of $\mathcal{H}$ satisfy (see p. 487 of [29])

$$N(\mathcal{H}, \eta) \leq (C/\eta)^n. \tag{4.8}$$

Combining this with (4.6), we obtain from (1.44)

$$\|f_\rho - f_{\mathbf{z}}\| \leq C_1 n^{-r} + \eta, \quad \mathbf{z} \in \Lambda_m(\eta), \tag{4.9}$$

where

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \leq 2(C_3/\eta^2)^n e^{-c_2 m\eta^2}. \tag{4.10}$$

The critical turning value in (4.10) occurs when $n[\ln C_3 + 2|\ln \eta|] = c_2 m\eta^2$. This gives

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \leq \begin{cases} Ce^{-cm\eta^2} & \eta \geq \eta_m \\ 1, & \eta \leq \eta_m, \end{cases} \tag{4.11}$$

where η_m as defined in the theorem. This proves (4.4). The estimate (4.5) follows by integrating (4.4) (see (1.19)). □

Let us mention some spaces W which satisfy the property (4.1). If $s > d/2$ and $p \geq 2$, then a theorem of Kashin can be used to deduce (4.1) for $W = W^s(L_p(X))$ where these Sobolev spaces are defined with respect to Lebesgue measure. Note that the assumption $s > d/2 \geq d/p$ guarantees that any ball $b(W^s(L_p(X))$ is compact in $\mathcal{C}(X)$. We therefore have the following corollary

Corollary 4.2 *If $W = W^s(L_p(X))$ with $s > d/2$ and $p \geq 2$, then the assumptions and conclusions of Theorem 4.1 hold for any ball $b(W)$.*

Figure 4.1 gives a graphical depiction of the smoothness spaces to which the Corollary applies. In the next section, we shall expand on this class of spaces by using nonlinear methods.

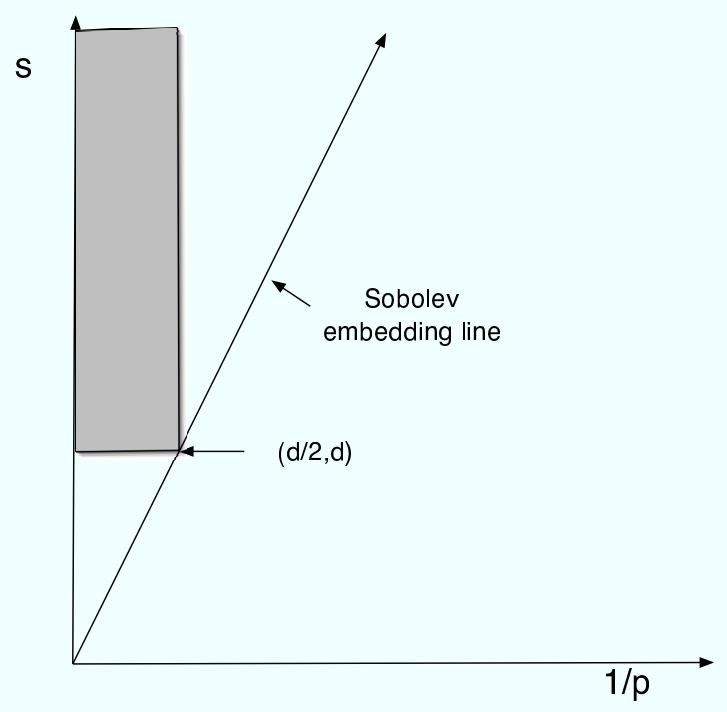

Figure 4.3: The grey shaded region indicates the smoothness spaces to which Corollary 4.2 apply.

4.2 Estimates based on nonlinear widths

We can improve upon the results of the previous subsection by using nonlinear widths in place of Kolmogorov widths. This will allow us to prove estimates like those in Theorem 4.1 but for a wider class of priors Θ.

We begin with the following setting for nonlinear widths given in [35]. Let $\mathcal{N}$ and n be positive integers. Given a Banach space $\mathcal{B}$, we shall look to approximate a given function $f \in \mathcal{B}$ using a collection $\Lambda_{\mathcal{N}} = \{\mathcal{L}_1, \ldots, \mathcal{L}_{\mathcal{N}}\}$ where each of the $\mathcal{L}_j$ are linear spaces of dimension n. This leads us to the following definition of $(\mathcal{N}, n)$-width for a compact class $K \subset \mathcal{B}$:

$$d_n(K, \mathcal{B}, \mathcal{N}) := \inf_{\Lambda_{\mathcal{N}}, \#\Lambda_{\mathcal{N}} \leq \mathcal{N}} \sup_{f \in K} \inf_{\mathcal{L} \in \Lambda_{\mathcal{N}}} \inf_{g \in \mathcal{L}} \|f - g\|_{\mathcal{B}}. \tag{4.12}$$

It is clear that

$$d_n(K, \mathcal{B}, 1) = d_n(K, \mathcal{B}). \tag{4.13}$$

The new feature of $d_n(K, \mathcal{B}, \mathcal{N})$ (as compared to $d_n(K, \mathcal{B})$) is that we have the ability to choose a subspace $\mathcal{L} \in \Lambda_{\mathcal{N}}$ depending on $f \in K$. It is clear that the bigger the value of $\mathcal{N}$, then the more flexibility we have to approximate f. It turns out that, from the point

of view of our applications, the following case

$$\mathcal{N} \asymp n^{an},$$

where $a > 0$ is a fixed number, plays an important role.

Let us assume that Θ is a compact subset of $\mathcal{C}(X)$ which satisfies $\Theta \subset b_{R_0}(\mathcal{C}(X))$, for some $R_0 > 0$ and also satisfies the following estimates for the nonlinear Kolmogorov widths

$$d_n(\Theta, \mathcal{C}(X), n^{an}) \le C_1 n^{-r}, \quad n = 1, 2, \dots. \tag{4.14}$$

Then by [35]

$$\epsilon_n(\Theta, \mathcal{C}(X)) \le C_2 (\ln n/n)^r, \quad n = 2, 3, \dots. \tag{4.15}$$

In the theorem that follows, we shall not be able to use Theorem C* directly since the set $\mathcal{H}$ we shall choose for the empirical least squares minimization will not be convex. Therefore, we first prove an extension of Theorem C* which deals with the nonconvex setting.

Theorem 4.3 *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that for all $f \in \mathcal{H}$, $f : X \to Y$ is such that $|f(x) - y| \le M$ a.e. Then, for all $\eta > 0$*

$$\rho^m\{\mathbf{z} : \|f_{\mathbf{z},\mathcal{H}} - f_{\mathcal{H}}\|^2 \ge \eta\} \le N(\mathcal{H}, \eta/(24M)) 2e^{-\frac{m\eta}{C(M,K)}} \tag{4.16}$$

provided $\|f_\rho - f_{\mathcal{H}}\|^2 \le K\eta$.

Proof The proof is similar to the proof of Theorem C* from [CS]. In the proof of Theorem C*, one uses the estimate (1.31) which we recall follows from the convexity assumption. In its place we shall use the estimate

$$\|f - f_{\mathcal{H}}\|^2 \le 2(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + 2K\eta), \quad f \in \mathcal{H}. \tag{4.17}$$

To prove this we note that

$$\begin{aligned} \|f - f_{\mathcal{H}}\|^2 &\le 2\{\|f - f_\rho\|^2 + \|f_{\mathcal{H}} - f_\rho\|^2\} = 2\{\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) \\ &+ \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) + \|f_{\mathcal{H}} - f_\rho\|^2\} = 2\{\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) + 2\|f_{\mathcal{H}} - f_\rho\|^2\}. \end{aligned} \tag{4.18}$$

Thus, (4.17) follows by placing our assumption $\|f_\rho - f_{\mathcal{H}}\|^2 \le K\eta$ into (4.18). The proof of Theorem 4.3 can now be completed in the same way as the proof of Theorem C*.

Theorem 4.4 *Let Θ satisfy (4.14). If $f_\rho \in \Theta$ and $m \in \{1, 2, \dots\}$, then there exists an estimator $f_{\mathbf{z}}$ such that*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \begin{cases} e^{-cm\eta^2} & \eta \ge \eta_m \\ 1, & \eta \le \eta_m, \end{cases} \tag{4.19}$$

where $\eta_m := C_2(\ln m/m)^{\frac{r}{1+2r}}$. In particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le C(\frac{\ln m}{m})^{\frac{r}{2r+1}} \tag{4.20}$$

where C is also an absolute constant.

Proof: The proof is very similar to that of Theorem 4.1. Given m, we shall choose $n := (\frac{m}{\ln m})^{\frac{1}{2r+1}}$. For this value of n let $\mathcal{N} := n^{an}$ with $a > 0$ given in (4.14). For this $\mathcal{N}$ and n there is a collection $\Lambda_{\mathcal{N}}$ of n-dimensional subspaces which realizes the approximation order (4.14). Here $\#(\Lambda_{\mathcal{N}}) = \mathcal{N}$. Thus for any $f \in b(W)$ there is an $\mathcal{L} \in \Lambda_{\mathcal{N}}$ and a $\phi_n \in \mathcal{L}$ such that $\|f - \phi_n\|_{\mathcal{C}(X)} \le C_1 n^{-r}$. It follows that $\|\phi_n\|_{\mathcal{C}(X)} \le R_0 + C_1 =: R$. We now consider the following set

$$\mathcal{H} := \cup_{\mathcal{L}\in\Lambda_{\mathcal{N}}} \mathcal{L} \cap b_R(\mathcal{C}(X)). \tag{4.21}$$

Then, it is clear that the entropy numbers for $\mathcal{H}$ satisfy

$$N(\mathcal{H}, \eta) \le \mathcal{N}(C/\epsilon)^n. \tag{4.22}$$

We define our estimator for $\mathbf{z} \in Z^m$ by

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

Using (4.22) with (4.14), we obtain from (4.16)

$$\|f_\rho - f_{\mathbf{z}}\| \le C_1 n^{-r} + \eta, \quad \mathbf{z} \in \Lambda_m(\eta), \tag{4.23}$$

where

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le 2\mathcal{N}(C_3/\eta^2)^n e^{-c_2 m\eta^2} \tag{4.24}$$

The critical turning value in (4.24) occurs when $an \ln n + n[\ln C_3 + 2|\ln \eta|] = c_2 m\eta^2$. This gives

$$\rho^m\{\mathbf{z} \notin \Lambda_m(\eta)\} \le \begin{cases} Ce^{-cm\eta^2} & \eta \ge \eta_{m,n} \\ 1, & \eta \le \eta_m, \end{cases} \tag{4.25}$$

with $\eta_{m,n} := C_2\sqrt{\frac{n\ln m}{m}}$ provided we choose C_2 large enough. This proves (4.19) and (4.20) follows by integrating (4.19). □

We give next an illustrative setting in which Theorem 4.4 can be applied. Let $\Psi := \{\psi_j\}_{j=1}^\infty$ be a Schauder basis for $\mathcal{C}(X)$. We fix an arbitrary $a > 0$ and consider for each positive integer n, the space Σ_{n,n^a} of all functions

$$\sum_{j\in\Gamma} c_j\psi_j, \quad \Gamma \subset \{1, \dots, n^a\}, \quad \#(\Gamma) \le n. \tag{4.26}$$

Thus we are in the same situation as in §2.5 except that we do not necessarily use an orthogonal system. As in §2.5, given any $f \in \mathcal{C}(X)$, we define

$$\sigma_{n,n^a}(f)_\infty := \inf_{g\in\Sigma_{n,n^a}} \|f - g\|_{\mathcal{C}(X)}. \tag{4.27}$$

This is the error of n-term approximation using Ψ except that we have imposed the extra condition on the indices.

We can realize this form of approximation as a special case of the approximation used in the definition of $(\mathcal{N}, n)$ widths with $\mathcal{N} := n^{an}$. Namely we consider the set of all n dimensional subspaces spanned by n elements of Ψ with the restriction that the indices of these elements come from $\{1, \dots, n^a\}$. There are $\le \mathcal{N}$ of these subspaces.

To carry this example further, we suppose that $X = I\!R^d$ and $\Psi := \{\psi_\lambda\}$ is a wavelet basis for $I\!R^d$ as described in §2.6. Approximation from Σ_{n,n^a} is then n-term wavelet approximation but with the restriction on the indices of basis functions. This form of approximation is used in encoding images and its approximation properties are well understood (see [9]).

Corollary 4.5 *Suppose that $f_\rho \in b(W)$ with $W = W^s(L_p(X)))$ with $s > d/p$ or that $W = B^s_q(L_p(X)$ with $s > d/p$ and $0 < q \le \infty$. Let $\mathcal{H} = \mathcal{H}_m := \Sigma_{n,n^a} \cap b_R(\mathcal{C}(X))$ be defined as above using the wavelet basis with $n := (\frac{m}{\ln m})^{\frac{d}{2s+d}}$ and $a > s - d/p$. Then, the least squares estimator $f_{\mathbf{z}}$ for this choice of $\mathcal{H}$ satisfies*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \begin{cases} e^{-cm\eta^2} & \eta \ge \eta_m \\ 1, & \eta \le \eta_m, \end{cases} \tag{4.28}$$

where $\eta_m := C_2(\ln m/m)^{\frac{s}{2s+d}}$. In particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le C(\frac{\ln m}{m})^{\frac{s}{2s+d}} \tag{4.29}$$

where C is also an absolute constant.

Proof: It was shown in [9] that for this choice of a, the above form of restricted approximation satisfies

$$\mathrm{dist}(f, \mathcal{H})_{\mathcal{C}(X)} \le C_1 n^{-\frac{s}{d}}. \tag{4.30}$$

Therefore, we can apply Theorem 4.4 and derive the Corollary. □

Notice that the Corollary applies to each smoothness space that is compactly embedded in $\mathcal{C}(X)$, i.e. to each smoothness space depicted in the shaded region of Figure 2.1.

4.3 Estimates for f_ρ using interpolation

We want to show in this section how techniques from the theory of interpolation of linear operators can be used to derive estimators $f_{\mathbf{z}}$ to f_ρ. The idea of using interpolation of operators was suggested in the paper of Smale and Zhou [32] in the setting of Hilbert spaces but they do not culminate this approach with concrete estimates since in the Hilbert space setting we do not have the analog of Theorem C*. We shall see that this approach falls a little short of giving the optimal decay ($O(m^{-\frac{s}{2s+d}})$) for Sobolev or Besov spaces of smoothness s.

We shall use interpolation with $\mathcal{C}(X)$ (equivalently $L_\infty(X)$) as one of the end point spaces. For the other end point space we can take $W_0 = W_0(X)$ where $W_0 \subset \mathcal{C}(X)$ is a smoothness space embedded in $\mathcal{C}(X)$. A space V is called an interpolation space for this pair $(\mathcal{C}(X), W_0)$ if each linear operator T which is bounded on both $\mathcal{C}(X)$ and W_0 is automatically bounded on V. The real method of interpolation gives one way to generate interpolation spaces by using what is called the K-functional:

$$K(f, t; \mathcal{C}(X), W_0) := \inf_{g \in W_0} \|f - g\|_{\mathcal{C}(X)} + t|g|_{W_0}. \tag{4.31}$$

We mention only one setting for this which will suffice for our analysis. Given $0 < \theta < 1$, we define $V_\theta := (\mathcal{C}(X), W_0)_{\theta,\infty}$ to be the set of all $f \in \mathcal{C}(X)$ such that

$$|f|_{V_\theta} := \sup_{t>0} t^{-\theta} K(f, t : \mathcal{C}(X), W_0) \tag{4.32}$$

is finite. So, membership in V_θ means that f can be approximated by a $g \in W_0$ to accuracy Ct^θ while the norm of g in W_0 is $\le Ct^{\theta-1}$:

$$\|f - g\|_{\mathcal{C}(X)} + t|g|_{W_0} \le |f|_{V_\theta} t^\theta. \tag{4.33}$$

We shall take for W_0 any Besov space $W_0 = B_p^s(L_p(X))$ which is compactly embedded in $\mathcal{C}(X)$. As mentioned earlier, we get a compact embedding if and only if $s > d/p$. It is known that the covering numbers for the unit ball $u(W_0)$ satisfy

$$N(\eta, u(W_0)) \le C_0 e^{c_0 \eta^{-d/s}}, \quad \eta > 0, \tag{4.34}$$

with the constants depending only on W_0.

Our main result of this section is the following.

Theorem 4.6 *Let W_0 be a Besov space $B_p^s(L_p(X))$ such that $u(W_0)$ is compactly embedded in $\mathcal{C}(X)$. If $f_\rho \in u(V_\theta)$ where $V_\theta = (\mathcal{C}(X), W_0)_{\theta,\infty}$ and $\theta := r/s$, then we take $\mathcal{H} := b_R(W_0)$ with $R := m^{\frac{s-r}{2r+d+rd/s}}$. The least squares minimizer $f_{\mathbf{z}}$ for this choice of $\mathcal{H}$ satisfies*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \begin{cases} e^{-cm\eta^2} & \eta \ge \eta_m \\ 1, & \eta \le \eta_m, \end{cases} \tag{4.35}$$

where $\eta_m := C_2 m^{-\frac{r}{2r+d+rd/s}}$. In particular,

$$E(\|f_\rho - f_{\mathbf{z}}\|) \le C m^{-\frac{r}{2r+d+rd/s}} \tag{4.36}$$

where C is a constant depending only on s and W_0.

Remark 4.7 *By rescaling, we can also treat the prior $f_\rho \in b_{R_0}(V_\theta)$ for any $R_0 > 0$.*

Proof: As usual, we only need to prove (4.35). We shall use the K-functional (4.31) but leave the choice of t open at the beginning. From the definition of V_θ we know that there is a function $g \in W_0$ such that

$$\|f_\rho - g\|_{\mathcal{C}(X)} + t|g|_{W_0} \le t^\theta |f_\rho|_{V_\theta} \le t^\theta. \tag{4.37}$$

Since $f_{\mathcal{H}}$ is a best approximation to f_ρ from $\mathcal{H}$ in the norm $\|\cdot\|$, it follows that the bias term satisfies

$$\|f_\rho - f_{\mathcal{H}}\| \le \|f_\rho - g\| \le \|f_\rho - g\|_{\mathcal{C}(X)} \le t^\theta. \tag{4.38}$$

The function g is in $b_{R_1}(W_0)$ where $R_1 = t^{\theta-1}$. Since $N(\eta, b_{R_1}(W_0)) = N(\eta/R_1, u(W_0))$, using (4.34) in (1.44) gives

$$\|f_\rho - f_{\mathbf{z}}\| \le t^\theta + \eta, \quad \mathbf{z} \in \Lambda_m \tag{4.39}$$

where

$$\rho^m\{\mathbf{z} \notin \Lambda_m\} \le e^{c_0(\frac{\eta^2}{R_1})^{-d/s} - c_1 m\eta^2}, \quad \eta > 0. \tag{4.40}$$

The turning value of η in (4.40) occurs when $\eta_* = cR_1^{\frac{d}{2s+2d}} m^{-\frac{s}{2s+2d}}$. Setting $t^\theta = \eta_*$ to balance the bias and variance gives

$$t = cm^{-\frac{s}{2r+d+rd/s}} \quad R_1 = cm^{\frac{s-r}{2r+d+rd/s}} \quad \eta_* = cm^{-\frac{r}{2r+d+rd/s}} \tag{4.41}$$

Since we do not know c, it is better, as stated in the Theorem, to use $R = m^{\frac{s-r}{2r+d+rd/s}}$ in place of R_1. This still leads to the estimate

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \begin{cases} e^{-cm\eta^2} & \eta \ge \eta_m \\ 1, & \eta \le \eta_m, \end{cases} \tag{4.42}$$

with η_m as stated in the Theorem. □

Any Besov space $B_q^r(L_p(X))$ which is compactly embedded in $\mathcal{C}(X)$ is contained in V_θ with $W_0 = B_\tau^s(L_\tau)X))$ with s arbitrarily large (see [8]). It follows from Theorem 4.6 that the estimates (4.35) and (4.36) hold for $f_\rho \in B_q^r(L_p)$ with s arbitrarily large. Still such estimates are not as good as those we have obtained in §4.2.

4.4 Universal estimators

We turn now to the problem of constructing universal estimators. As a starting point, recall the analysis of §4.1 of linear estimators. If we have a prior class Θ and we know the parameter r of its approximation order, then we choose our estimator from the linear space of dimension $n := (\frac{m}{\ln m})^{\frac{1}{2r+1}}$. Our goal now is to construct an estimator which does not need to know r but is simultaneously optimal for all possible values of r. There is a common technique in statistics, known as penalty methods for constructing such estimators (see e.g. Chapter 12 of [19], see also [4] and [38]). The point of this section is to analyze the performance of one such penalty method. In the first part of this section, we shall bound the accuracy of this estimator in probability. Unfortunately, to accomplish this we shall impose rather stringent assumptions on the parameter r; namely that $r \le 1/2$. It would be of great interest to remove this restriction on r. In the second part of this section, we shall consider bounds on our estimator in expectation rather than probability. This will enable us to remove the restriction $r \le 1/2$. We should also mention that universal estimators are given in [26] and also in [5] using a completely different technique. The advantage of the estimators given in [5] is that they do not go through L_∞ and thereby apply to weaker smoothness conditions imposed on f_ρ. They also have certain numerical advantages.

We shall put ourselves in the following setting. We suppose that we have in hand a sequence $(\mathcal{L}_n)$ of linear subspaces of $\mathcal{C}(X)$ with $\mathcal{L}_n$ of dimension n. For each $r > 0$, we denote by W^r a normed linear space of functions such that

$$\operatorname{dist}(u(W^r), \mathcal{L}_n)_{\mathcal{C}(X)} := \sup_{f \in u(W^r)} \inf_{g \in \mathcal{L}_n} \|f - g\|_{\mathcal{C}(X)} \le C_0 n^{-r}, \quad n = 1, 2, \dots, \tag{4.43}$$

with C_0 an absolute constant and with $u(W^r)$ denoting, as usual, the unit ball of this space. Thus we are in a setting similar to our treatment of Kolmogorov's n-widths. An example will be given at the end of this section.

We want to give an estimator $f_{\mathbf{z}}$ which will approximate f_ρ whenever f_ρ is in any of the $u(W^r)$. However, the estimator should work without knowledge of r. As in the discussion of estimators based on Kolmogorov's widths, we know there is an R depending only on C_0 such that for $\mathcal{H}_n := \mathcal{L}_n \cap b_R(\mathcal{C}(X))$, we have

$$\operatorname{dist}(u(W^r), \mathcal{H}_n)_{\mathcal{C}(X)} := \sup_{f \in u(W^r)} \inf_{g \in \mathcal{H}_n} \|f - g\|_{\mathcal{C}(X)} \le C_0 n^{-r}, \quad n = 1, 2, \dots. \tag{4.44}$$

We define the estimator $f_{\mathbf{z}}$ by the formula

$$f_{\mathbf{z}} := f_{\mathbf{z},\mathcal{H}_k} \tag{4.45}$$

with

$$k := k(\mathbf{z}) := \arg\min_{1 \le j \le m} (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_j}) + \frac{Aj \ln m}{m}) \tag{4.46}$$

where $A > 1$ is a constant whose exact value will be spelled out below.

We want to analyze how well $f_{\mathbf{z}}$ approximates f_ρ. For this we shall use the following lemma.

Lemma 4.8 *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$ and let $\epsilon > 0$. Then for all $f \in \mathcal{H}$*

$$\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) \le 2(\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})) + 2\epsilon. \tag{4.47}$$

holds for all $\mathbf{z} \notin \Lambda(\mathcal{H}, \epsilon)$ where

$$\rho^m \Lambda(\mathcal{H}, \epsilon) \le N(\mathcal{H}, \frac{\epsilon}{24M}) \exp(-\frac{m\epsilon}{288M^2}). \tag{4.48}$$

Proof: This is an immediate consequence of Proposition 7 in [10] with α chosen to be $1/6$ in that Proposition. □

Remark 4.9 *If like in Lemma 4.8, we assume $|f(x) - y| \le M$, for all $(x, y) \in Z$ and all $f \in \mathcal{H}$, then we can drop the assumption of convexity in the Lemma and draw the same conclusion with $f_{\mathcal{H}}$ replaced by f_ρ. This can be proved in the same way as Lemma 4.8 (see [10]) and it also can be derived from Theorem 11.4 in [19] (with different constants).*

Theorem 4.10 *Let $f_{\mathbf{z}}$ be defined by (4.45). There are suitably chosen constants $C, A \ge 1$ and $c > 0$ such that whenever $f_\rho \in u(W^r)$, for some $r \in [a, 1/2]$ then for all $m \ge 3$,*

$$\rho^m\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\| \ge \eta\} \le C \begin{cases} e^{-cm\eta^4} & \eta \ge \eta_{m,r} \\ 1, & \eta \le \eta_{m,r}, \end{cases} \tag{4.49}$$

where $\eta_{m,r} := \sqrt{A}(\ln m/m)^{\frac{r}{2r+1}}$. In particular,

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|) \le C(\frac{\ln m}{m})^{\frac{r}{2r+1}} \tag{4.50}$$

where C is again an absolute constant.

Remark 4.11 *Notice that when $\eta \geq \eta_{m,r}$, then $m\eta^4 \geq m\eta_{m,r}^4 \geq A^2(\ln m)^{\frac{4r}{2r+1}} m^{1-\frac{4r}{2r+1}} \geq A^2(\ln m)$ because $\ln m \leq m$ and $r \leq 1/2$. In particular, $e^{-m\eta_{m,r}^4}$ tends to zero as $m \to \infty$*

Proof: The estimate (4.50) follows from (4.49). To prove (4.49), we fix $r \in [a, 1/2]$ and assume that $f_\rho \in u(W^r)$. We note that we have nothing to prove when $\eta \leq \eta_{m,r}$. Also, we have nothing to prove if $\eta > R + M$ because$\|f_\rho\| \leq M$ and $\|f_{\mathbf{z}}\| \leq R$. Also, the estimate for $1 < \eta \leq M + R$ will follow from the estimate for $\eta = 1$ (with an adjustment in constants). Therefore, in going further, we assume that $\eta_{m,r} \leq \eta \leq 1$.

Let us begin by applying Bernstein's inequality to the random variable $(y - f_{\mathcal{H}_j}(x))^2$ and find for any such η:

$$|\mathcal{E}(f_{\mathcal{H}_j}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_j})| \leq \eta^2, \quad \mathbf{z} \notin \Lambda_1(\eta, j), \tag{4.51}$$

where

$$\rho^m \Lambda_1(\eta, j) \leq 2e^{-c_1 m\eta^4} \tag{4.52}$$

where $c_1 > 0$ depends only on M and R. We define $\Lambda_1(\eta) := \cup_{j=1}^m \Lambda_1(\eta, j)$. Then, with a view towards Remark 4.11, we see that

$$\rho^m \Lambda_1(\eta) \leq 2me^{-c_1 m\eta^4} \leq e^{\ln(2m) - c_1 m\eta^4} \leq e^{-cm\eta^4} \tag{4.53}$$

provided $A^2 > 2/c_1$ which is our first requirement on A.

Let us now define n as the smallest integer such that $n\frac{\ln m}{m} \geq \eta^2$. Notice that $n \geq 2$ because $\eta \geq \eta_{m,r}$. For each $1 \leq j \leq m$, we define

$$\epsilon_j := A \begin{cases} \eta^2, & 1 \leq j \leq n, \\ \frac{j \ln m}{m}, & n < j \leq m. \end{cases} \tag{4.54}$$

and define $\Lambda_2(\eta) := \cup_{j=1}^m \Lambda(\mathcal{H}_j, \epsilon_j)$ where the sets $\Lambda(\mathcal{H}_j, \epsilon_j)$ are those appearing in Lemma 4.8. From (4.8), we have $N(\mathcal{H}_j, \epsilon_j/24M) \leq (C_2/\epsilon_j)^j$ for some constant $C_2 > 0$. Hence,

$$\rho^m \Lambda_2(\eta) \leq \sum_{1 \leq j \leq n} e^{-j(\ln \epsilon_j - \ln C_2) - c_2 m\epsilon_j} + \sum_{n < j \leq m} e^{-j(\ln \epsilon_j - \ln C_2) - c_2 m\epsilon_j} = \Sigma_1 + \Sigma_2, \tag{4.55}$$

with $c_2 = (288M^2)^{-1}$. We shall require that $A \geq C_2$ and $A \geq 4/c_2$. This finishes the conditions on A and we now fix A as the smallest number satisfying the three requirements we have stipulated.

With this choice of A, it follows that the exponent of each summand in Σ_1 is $\leq -2j \ln \eta - c_2 mA\eta^2$. This means that this sum is bounded by a geometric series dominated by the term $j = n$. Since

$$-2n \ln \eta \leq n \ln(\frac{m}{\ln m}) \leq n \ln m \leq 2m\eta^2 < c_2 Am\eta^2. \tag{4.56}$$

This means that $\Sigma_1 \leq e^{-cm\eta^2}$ for an absolute constant $c > 0$.

We can use similar reasoning to derive the same bound for Σ_2. Namely, the exponent of each summand in Σ_2 does not exceed

$$-j(\ln \epsilon_j - \ln C_2) - c_2 m\epsilon_j \leq j \ln m - Ac_2 j \ln m \leq -Ac_2 j \ln m/2. \tag{4.57}$$

So this sum is also bounded by a geometric series whose sum is in turn dominated by $Ce^{-Ac_2 n \ln m/2} \leq e^{-cm\eta^2}$ with c another constant.

In summary, we have shown that

$$\rho^m(\Lambda_1(\eta) \cup \Lambda_2(\eta)) \leq e^{-cm\eta^4} + e^{-cm\eta^2} \leq e^{-cm\eta^4} \tag{4.58}$$

for some constant $c > 0$.

Going further, we shall only consider $\mathbf{z} \notin \Lambda_1(\eta) \cup \Lambda_2(\eta)$. For any such $\mathbf{z}$, we have from (4.47)

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z}}) &= \mathcal{E}(f_{\mathbf{z},\mathcal{H}_k}) \leq 2\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}_k}) + 2(\mathcal{E}(f_{\mathcal{H}_k}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_k})) + 2\epsilon_k \\ &\leq 2\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}_k}) + 2\eta^2 + 2\epsilon_k \leq 2\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + 2\eta^2 + 2\epsilon_k, \end{aligned} \tag{4.59}$$

where we used (4.51) and the fact that $\mathcal{E}(f_\rho) \leq \mathcal{E}(f_{\mathcal{H}_k})$.

From the definition of k, we have (note that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_n}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_n})$)

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_n}) + A\frac{(n-k)\ln m}{m} \leq \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_n}) + 2A\eta^2 - \frac{Ak\ln m}{m}. \tag{4.60}$$

Therefore, returning to (4.59) we derive

$$\begin{aligned} \|f_\rho - f_{\mathbf{z}}\|^2 &= \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \\ &\leq 2(\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_n}) - \mathcal{E}(f_\rho)) + (4A+2)\eta^2 + 2\epsilon_k - 2A\frac{k\ln m}{m} \\ &\leq 2(\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_n}) - \mathcal{E}(f_\rho)) + (6A+2)\eta^2 \end{aligned} \tag{4.61}$$

where the last inequality follows from the definition of ϵ_k. Since $\mathbf{z} \notin \Lambda_1(\eta)$, we can replace $\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}_n})$ by $\mathcal{E}(f_{\mathcal{H}_n})$ and in doing so we incur an error of most η^2. This gives

$$\begin{aligned} \|f_\rho - f_{\mathbf{z}}\|^2 &\leq 2(\mathcal{E}(f_{\mathcal{H}_n}) - \mathcal{E}(f_\rho)) + (6A+4)\eta^2 \\ &= 2\|f_\rho - f_{\mathcal{H}_n}\|^2 + (6A+4)\eta^2 \\ &\leq 2C_0^2 n^{-2r} + (6A+4)\eta^2 \\ &\leq (2C_0^2 + 6A + 4)\eta^2. \end{aligned} \tag{4.62}$$

Here in bounding $C_0^2 n^{-2r}$, we have used the fact that

$$n \geq \eta^2 \frac{m}{\ln m} \geq A\left(\frac{m}{\ln m}\right)^{\frac{1}{2r+1}} \geq A^{\frac{2r+1}{2r}}\eta^{-1/r} \geq \eta^{-1/r} \tag{4.63}$$

where the first inequality follows from the definition of n and the next two from the restriction $\eta \geq \eta_{m,r}$. The theorem now follows easily from (4.62) together with (4.58). □

We shall next consider bounds in expectation for the estimator (4.45). In this setting, we shall be able to replace the assumption that $a \leq r \leq 1/2$ by $a \leq r \leq b$ for any $b > 0$.

Theorem 4.12 *Let $f_{\mathbf{z}}$ be defined by (4.45) with $A \geq 1$ chosen sufficiently large. If $f_\rho \in u(W^r)$, for some $r > 0$, then for all $m \geq 3$,*

$$E_{\rho^m}(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C(r)\left(\frac{\ln m}{m}\right)^{\frac{2r}{2r+1}} \tag{4.64}$$

where $C(r)$ is bounded on any interval $[a,b]$ with $0 < a < b < \infty$.

Proof: Let $k = k(\mathbf{z})$ be as in 4.46 and let $\epsilon_j := \frac{Aj \ln m}{m}$ for each $j = 1, 2, \ldots, m$. Throughout this proof the expectation E is with respect to ρ^m. For the set $\Lambda(\mathcal{H}_k, \epsilon_k)$ given by Lemma 4.8,

$$\begin{aligned} E(\|f_\rho - f_{\mathbf{z}}\|^2) &= E(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)) \\ &= \int\limits_{\Lambda(\mathcal{H}_k,\epsilon_k)} (\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho))\, d\rho^m + \int\limits_{Z^m \setminus \Lambda(\mathcal{H}_k,\epsilon_k)} (\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho))\, d\rho^m \\ &= I_1 + I_2. \end{aligned} \tag{4.65}$$

Using the boundedness of $f_{\mathbf{z}}$ and f_ρ, we obtain from Remark 4.9,

$$I_1 \le C\rho^m(\Lambda(\mathcal{H}_k, \epsilon_k) \le CN(\mathcal{H}_k, \epsilon_k/24M)e^{-\frac{m\epsilon_k}{288M^2}} \le C(C_2/\epsilon_k)^k e^{-\frac{Ak\ln m}{288M^2}} \le Cm^{-1} \tag{4.66}$$

provided we take A sufficiently large.

To estimate I_2, we again use Remark 4.9 and find

$$I_2 \le 2 \int\limits_{Z^m \setminus \Lambda(\mathcal{H}_k,\epsilon_k)} (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho) + \epsilon_k)\, d\rho^m \le 2E(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho) + \epsilon_k)). \tag{4.67}$$

Now notice that

$$E((\mathcal{E}_{\mathbf{z}}(f_\rho)) = \frac{1}{m}\sum_{i=1}^m \int_{Z^m} (f_\rho(x_i) - y_i)^2 d\rho^m = \mathcal{E}(f_\rho). \tag{4.68}$$

Also, by the definition of k and $f_{\mathbf{z}}$, we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \epsilon_k = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_k}) + \epsilon_k = \min_{1\le j\le m} (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_j}) + \epsilon_j) \tag{4.69}$$

Therefore,

$$E(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \epsilon_k) \le \min_{1\le j\le m} (E(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_j})) + \epsilon_j). \tag{4.70}$$

Since by the definition of $f_{\mathbf{z},\mathcal{H}_j}$, we have $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_j}) = \inf_{f\in\mathcal{H}_j} \mathcal{E}_{\mathbf{z}}(f)$, it follows that

$$E(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{H}_j})) \le \inf_{f\in\mathcal{H}_j} E(\mathcal{E}_{\mathbf{z}}(f)) = \inf_{f\in\mathcal{H}_j} \mathcal{E}(f). \tag{4.71}$$

To complete our estimate of I_2, we use the definition of W^r and obtain

$$\inf_{f\in\mathcal{H}_j} \mathcal{E}(f) - \mathcal{E}(f_\rho) = \inf_{f\in\mathcal{H}_j} \|f - f_\rho\|^2 \le C_1^2 j^{-2r}. \tag{4.72}$$

Combining (4.68), (4.70), and (4.71), we obtain

$$E(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \epsilon_k) \le \min_{1\le j\le m} \left(C_1^2 j^{-2r} + \frac{Aj\ln m}{m} \right) + \mathcal{E}(f_\rho) \le C\left(\frac{\ln m}{m}\right)^{\frac{2r}{2r+1}} + \mathcal{E}(f_\rho), \tag{4.73}$$

where the last inequality was obtained by choosing j as close to $(\frac{m}{\ln m})^{\frac{1}{2r+1}}$ as possible. Substituting (4.68) and (4.73) into (4.67), we obtain $I_2 \le C(\frac{\ln m}{m})^{\frac{2r}{2r+1}}$. When this estimate is combined with (4.66) we complete the proof of the Theorem. □

It is quite straight forward to extend Theorem 4.12 to apply to nonlinear methods as described in §4.2. Instead of using linear spaces, for each n, we define $N(n) := \lceil n^{an} \rceil$ and now take a collection $\Lambda_n := \{\mathcal{L}_j(n)\}_{j=1}^{N(n)}$ of linear spaces $\mathcal{L}_j(n)$, each of dimension n. In place of W^r, $r > 0$, we use the class $W^r(\{\Lambda_n\})$ which is defined as the set of all f such that $\|f\|_{\mathcal{C}(X)} \le R_0$ and

$$\inf_{1\le j\le N(n)} \operatorname{dist}(f, \mathcal{L}_j(n))_{\mathcal{C}(X)} \le C_0 n^{-r}, \quad n = 1, 2, \dots. \tag{4.74}$$

As our hypothesis class, we take $\mathcal{H}_n := \cup_{j=1}^{N(n)} (\mathcal{L}_j(n) \cap b_R(\mathcal{C}(X)))$ with $R := R_0 + C_0$. Then we define $f_{\mathbf{z}}$ by the formula (4.45) with this choice for the $\mathcal{H}_n$. We obtain that Theorem 4.12 now holds with these choices and the same proof.

Let us mention an example of how Theorem 4.10 can be applied. We consider the Sobolev spaces $W^s(\mathcal{C}(X))$ with $X = [0,1]^d$ and $a \le s \le d/2$. We can take for $\mathcal{L}_n$ one of several classical approximation spaces. For example, we could use $\mathcal{L}_n$ to be an n-dimensional space spanned by the first n wavelets from a wavelet orthogonal system or we could take piecewise polynomials of degree $\ge d/2$ on a uniform subdivision of X into cubes. It is well known that in either of these two settings, we have

$$\operatorname{dist}(u(W^s(\mathcal{C}(X)), \mathcal{L}_n)_{\mathcal{C}(X)}) \le C n^{-s/d}. \tag{4.75}$$

Therefore, Theorem 4.10 applies and we have a universal estimator for this family of Sobolev spaces. When we seek estimates in expectation as in Theorem 4.12, we can remove the restriction that $s \le d/2$.

By using nonlinear methods of approximation we can widen the applicability of Theorem 4.12 to any Besov space which compactly embeds into $\mathcal{C}(X)$. Here for Λ_n, we take the wavelet system as described in §4.2 which corresponds to n-term wavelet approximation. Namely, if $p > d/s$ and $0 < q \le \infty$ then for any ball Θ in the Besov space $B_q^s(L_p(X))$, Theorem 4.12 holds with $r = s/d$.

5 A variant of the regression problem

In this section, we shall treat a variant of the regression problem. We shall now assume that X is a cube in $\mathbb{R}^d$. Without loss of generality we can take $X = [0,1]^d$. We will also assume that ρ_X is an absolutely continuous measure with density $\mu(x)$, that is, $d\rho_X = \mu dx$. We continue to assume that $|y| \le M$. Thus, we are slightly more restrictive than earlier where we had no restrictions on ρ_X.

In place of estimating the regression function f_ρ, we shall instead estimate the function

$$f_\mu := \mu f_\rho \tag{5.1}$$

in one of the L_p norms (quisi-norms)

$$\|g\|_{L_p} := \left(\int_X |g(x)|^p \, dx \right)^{1/p}, \quad 0 < p < \infty \tag{5.2}$$

where now these norms are taken with respect to Lebesgue measure.

In order to explain our original motivation for estimating f_μ, we return to the problem of bank loans that was described in the introduction. A possible goal of the bank is to find $\Omega \subset X$ that maximizes

$$\int_\Omega f_\rho(x) d\rho_X$$

since this is related to maximizing profit. The optimal choice for such Ω is given by

$$\Omega_0 = \{x : f_\rho(x) \geq 0\}.$$

The mathematical question in this regard is how to utilize the available data $\mathbf{z}$ to find an empirical $\Omega_\mathbf{z}$ that gives a good approximation to Ω_0.

We suggest the following way to solve this problem. From the definition of f_μ we have

$$\int_\Omega f_\rho(x) d\rho_X = \int_\Omega f_\mu(x) dx.$$

Suppose we have found $f_\mathbf{z}$ such that

$$\rho^m\{\mathbf{z} : \|f_\mu - f_\mathbf{z}\|_{L_1} \geq \epsilon\} \leq \delta.$$

Define

$$\Omega_\mathbf{z} := \{x : f_\mathbf{z}(x) \geq 0\}.$$

Then with the above estimate on the probability we have

$$\int_{\Omega_\mathbf{z}} f_\rho(x) d\rho_X = \int_{\Omega_\mathbf{z}} f_\mu(x) dx \geq \int_{\Omega_\mathbf{z}} f_\mathbf{z}(x) dx - \epsilon$$

$$\geq \int_{\Omega_0} f_\mathbf{z}(x) dx - \epsilon \geq \int_{\Omega_0} f_\rho(x) d\rho_X - 2\epsilon.$$

Therefore, the empirical set $\Omega_\mathbf{z}$ provides an optimal profit within an error 2ϵ with probability $\geq 1 - \delta$.

We shall construct estimators for f_μ based on linear approximation from orthogonal systems and prove they are semi-optimal for certain smoothness space priors. In classical settings, we can take either Fourier or wavelet orthogonal systems. We shall measure the error in L_2 but note that a similar analysis could be applied for L_p estimation. We will not have to go through L_∞ to derive our estimates as we did in the treatment of f_ρ. Therefore, the range of smoothness conditions that apply to f_μ correspond to any smoothness space compactly embedded in L_2. In our estimates for f_ρ , the embedding had to be into L_∞ even though we were measuring discrepancy with respect to $L_2(X, \rho_X)$

5.1 Estimators based on orthogonal systems.

In this section, we shall construct linear estimators based on orthogonal systems. Let $\{\psi_j\}_{j=1}^{\infty}$ be an orthonormal system for $L_2(X)$ with respect to Lebesgue measure. We assume that $f_\mu \in b(W_r)$ where this class has the property that $g \in b(W_r)$ implies

$$\|g - S_j(g)\|_{L_2} \le C_0 j^{-r}, \quad j = 1, 2, \ldots, \tag{5.3}$$

where S_j is given by (2.21). We have already mentioned that in the case of the Fourier or wavelet bases, we can take $W_r := B_\infty^{rd}(L_2(X))$ where the latter is the Besov space on X.

We assume that r is known to us. Given m, we define $n := \lfloor \frac{m}{C_1 \ln m} \rfloor^{\frac{1}{2r+1}}$ with $C_1 > 1$ specified below. Our estimator is

$$f_{\mathbf{z}} := \sum_{j=1}^{n} \hat{c}_j(\mathbf{z})\psi_j, \tag{5.4}$$

where

$$\hat{c}_j(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^{m} y_i \psi_j(x_i), \quad j = 1, 2, \ldots, n. \tag{5.5}$$

We shall first deal with the case that the bases functions ψ_j are uniformly bounded $\|\psi_j\|_{L_\infty} \le C_2$. In this case, the random variable $y\psi_j(x)$ has variance $\le C_2^2 M^2$ and L_∞ norm $\le C_2 M$, and Bernstein's inequality gives

$$\rho^m\{\mathbf{z} : |\hat{c}_j(\mathbf{z}) - c_j| \ge \eta\} \le 2\epsilon^{-\frac{m\eta^2}{2(C_2^2M^2 + C_2 M\eta/3)}}, \tag{5.6}$$

for each $\eta > 0$.

Theorem 5.1 *Suppose that the basis functions ψ_j are uniformly bounded by C_2. If $f_\mu \in b(W_r)$, $r > 0$, then whenever the constant C_1 is chosen sufficiently large, the estimator $f_{\mathbf{z}}$ defined by (5.4) satisfies*

$$\rho^m\{\mathbf{z} : \|f_\mu - f_{\mathbf{z}}\|_{L_2} \ge \eta\} \le \begin{cases} 1, & \eta \le \eta_m, \\ e^{-\frac{cm\eta^2}{n}}, & \eta_m \le \eta \le 1/\sqrt{n}, \\ e^{-\frac{cm\eta}{\sqrt{n}}}, & \eta > 1/\sqrt{n}, \end{cases} \tag{5.7}$$

where $\eta_m := (C_1 \ln m/m)^{\frac{r}{2r+1}}$. In particular,

$$E(\|f_\mu - f_{\mathbf{z}}\|_{L_2}) \le C(\frac{\ln m}{m})^{\frac{r}{2r+1}} \tag{5.8}$$

where C is an absolute constant.

Proof: The estimate (5.8) follows from (5.7) (see (1.19)). Therefore, we concentrate on proving (5.7). We can assume that $\eta \ge \eta_m$. We write $f_\mu - f_{\mathbf{z}} = f_\mu - S_n(f_\mu) + S_n(f_\mu) - f_{\mathbf{z}}$. The L_2 norm of the first term is bounded by $C_0 n^{-r}$ (see (5.3). Thus we have

$$\|f_\mu - f_{\mathbf{z}}\|_{L_2} \le C_0 n^{-r} + \left(\sum_{j=1}^{n} |\hat{c}_j(\mathbf{z}) - c_j|^2\right)^{1/2}. \tag{5.9}$$

Given $\eta > 0$, we define $\Lambda_j(\eta) := \{\mathbf{z} : |c_j - \hat{c}_j(\mathbf{z})| \geq \eta/\sqrt{n}\}$ and $\Lambda(\eta) := \cup_{j=1}^n \Lambda_j(\eta)$. For $\mathbf{z} \notin \Lambda(\eta)$ we have from (5.9)

$$\|f_\mu - f_{\mathbf{z}}\|_{L_2} \leq C_0 n^{-r} + \eta \leq (C_0 + 1)\eta. \tag{5.10}$$

From (5.6), we know that for $\eta_m \leq \eta \leq 1/\sqrt{n}$,

$$\rho^m\{\mathbf{z} \in \Lambda(\eta)\} \leq 2ne^{-c_1 \frac{m\eta^2}{n}} \leq e^{-\frac{cm\eta^2}{n}}, \tag{5.11}$$

where in the last inequality we used the fact that $c_1 m\eta_m^2/n \geq c_1 C_1(\ln m)$ to absorb the factor $2n$ into the exponent by an appropriate choice of c. This can be done provided $c_1 C_1 \geq 3$ which is a condition we impose on C_1. When $\eta > 1/\sqrt{n}$, we have

$$\rho^m\{\mathbf{z} \in \Lambda(\eta)\} \leq 2ne^{-c_2 \frac{m\eta}{\sqrt{n}}} \leq e^{-c\frac{m\eta}{\sqrt{n}}} \tag{5.12}$$

where we again absorb the factor $2m$ into the exponential. From these two probability estimates and (5.10), we easily complete the proof of the theorem. □

We shall next show how to modify the above ideas to give a similar result in the case of the wavelet basis. We shall use the notation ψ_I^e, $I \in \mathcal{D}_+$, $e \in E$, which was given in §2.6. Recall that ψ_I^e is supported on a cube $\tilde{I}$ which is a fixed expansion of I. At a given dyadic level j, any point $x \in X$ is in at most C_3 cubes $\tilde{I}$, $I \in \mathcal{D}_j$ and therefore

$$\sum_{I \in \mathcal{D}_j} \chi_{\tilde{I}}(x) \leq C_3. \tag{5.13}$$

For each basis function ψ_I^e, we have that the random variable $y\psi_I^e(x)$ satisfies

$$\|y\psi_I^e(x)\|_{L_\infty} \leq C_2 M |I|^{-1/2} \quad \text{and} \quad \sigma^2(y\psi_I^e(x)) \leq C_2^2 M^2 |I|^{-1}\rho(\tilde{I}). \tag{5.14}$$

It follows therefore from Bernstein's inequality applied to this random variable that for any of the first n coefficients c_I^e we have

$$\rho^m\{\mathbf{z} : |\hat{c}_I^e(\mathbf{z}) - c_I^e| \geq \epsilon\} \leq 2e^{-\frac{m\epsilon^2}{2(C_2^2 M^2 n\rho(\tilde{I}) + C_2 M\sqrt{n}\epsilon/3)}}, \tag{5.15}$$

for each $\eta > 0$.

As before, we denote by $b(W_r)$ a class of functions g that satisfy (5.3). Given m, we define $n := \lceil \frac{m}{C_1 \ln m} \rceil^{\frac{1}{2r+1}}$ and the estimator

$$f_{\mathbf{z}} := \sum_{(I,e) \in \Gamma_n} \hat{c}_I^e(\mathbf{z})\psi_I^e \tag{5.16}$$

where Γ_n is the set of indices corresponding to the first n wavelets and the $\hat{c}_j(\mathbf{z})$ are defined in (5.5).

Theorem 5.2 *Suppose that $\{\psi_I^e\}$ is a wavelet basis for $[0,1]^d$. If $f_\mu \in b(W_r)$, $r > 0$, then whenever the constant C_2 is chosen sufficiently large, the estimator $f_{\mathbf{z}}$ defined by (5.16) satisfies*

$$\rho^m\{\mathbf{z} : \|f_\mu - f_{\mathbf{z}}\|_{L_2} \geq \eta\sqrt{\ln m}\} \leq \begin{cases} 1, & \eta \leq \eta_m, \\ e^{-\frac{cm\eta \min(\eta,1)}{n}}, & \eta_m \leq \eta, \end{cases} \tag{5.17}$$

for $m = 3, 4, \ldots$, where $\eta_m := (C_1 \ln m/m)^{\frac{r}{1+2r}}$. In particular,

$$E(\|f_\mu - f_{\mathbf{z}}\|_{L_2}) \leq C\sqrt{\ln m}(\frac{\ln m}{m})^{\frac{r}{2r+1}} \tag{5.18}$$

where C is an absolute constant.

Proof: As in Theorem 5.1, we only have to prove (5.17) for $\eta \geq \eta_m$ since the rest of the theorem follows easily from this. We define $\lambda_I := \max(\rho(\tilde{I}), n^{-1})$. Given η, we define

$$\Lambda_I^e(\eta) := \{\mathbf{z} : |c_I^e - \hat{c}_I^e(\mathbf{z})| \geq \eta\sqrt{\lambda_I}\}, \quad (I, e) \in \Gamma_n \tag{5.19}$$

and $\Lambda(\eta) := \cup_{(I,e)\in\Gamma_n}\Lambda_I^e(\eta)$. Further, we define Γ_n^+ to be the set of those indices $(I, e) \in \Gamma_n$ for which $\lambda_I = \rho(\tilde{I})$ and $\Gamma_n^- := \Gamma_n \setminus \Gamma_n^+$. Then, whenever $\mathbf{z} \notin \Lambda(\eta)$ and $(I, e) \in \Gamma_n^-$, we have $|c_I^e - \hat{c}_I^e(\mathbf{z})| \leq \eta/\sqrt{n}$,and therefore

$$\sum_{(I,e)\in\Gamma_n^-} |\hat{c}_I^e(\mathbf{z}) - c_I^e|^2 \leq \eta^2. \tag{5.20}$$

On Γ_n^+ we have $|\hat{c}_I^e(\mathbf{z}) - c_I^e| \leq \eta\sqrt{\rho(\tilde{I})}$ whenever $\mathbf{z} \notin \Lambda(\eta)$. Let $\Gamma_n^+(j) := \Gamma_n^+ \cap \mathcal{D}_j$ be the collection of those indices corresponding to dyadic level j. Then,

$$\sum_{(I,e)\in\Gamma_n^+(j)} |\hat{c}_I^e(\mathbf{z}) - c_I^e|^2 \leq C_3\eta^2, \tag{5.21}$$

where we have used the overlapping property (5.13) and the fact that $\rho_X(X) = 1$. Note that there are at most $C \ln n$ dyadic levels active in S_n. Therefore, summing over all these dyadic levels we obtain

$$\sum_{(I,e)\in\Gamma_n} |\hat{c}_I^e(\mathbf{z}) - c_I^e|^2 \leq (1 + C_3 \ln n)\eta^2. \tag{5.22}$$

This leads to the estimate

$$\|f_\mu - f_{\mathbf{z}}\|_{L_2} \leq C\eta(\sqrt{\ln m}) \tag{5.23}$$

with $C > 0$ an absolute constant. This is the estimate we want for the error.

Now, we estimate the probability that $\mathbf{z} \in \Lambda(\eta)$. Looking at (5.15), the first term in the denominator dominates when $\eta \leq C_4\sqrt{n}\rho(\tilde{I})/\sqrt{\lambda_I}$ with C_4 a fixed constant. Therefore, we obtain

$$\rho^m\{\mathbf{z} \in \Lambda_I^e(\eta)\} \leq \begin{cases} e^{-\frac{c_1 m\eta^2\lambda_I}{n\rho_I}}, & \eta \leq C_4\sqrt{n}\rho(\tilde{I})/\sqrt{\lambda_I}, \\ e^{-\frac{c_1 m\eta\sqrt{\lambda_I}}{\sqrt{n}}}, & \eta > C_4\sqrt{n}\rho(\tilde{I})/\sqrt{\lambda_I}, \end{cases} \tag{5.24}$$

In other words,

$$\rho^m\{\mathbf{z} \in \Lambda_I^e(\eta)\} \leq e^{-c_1\frac{m\eta}{n}\min(\frac{\eta\lambda_I}{\rho_I}, \sqrt{\lambda_I n})} \leq e^{-c_1\frac{m\eta}{n}\min(\eta,1)} \tag{5.25}$$

This gives that for $\eta \geq \eta_m$,

$$\rho^m\{\mathbf{z} \in \Lambda(\eta)\} \leq ne^{-c_1\frac{m\eta}{n}\min(\eta,1)} \leq e^{-c\frac{m\eta}{n}\min(\eta,1)} \tag{5.26}$$

where we have absorbed the factor n into the exponential in the usual way.

From these two probability estimates and (5.23), we easily complete the proof of the theorem. □

References

[1] P.S. Alexandroff, *Combinatorial Topology*, Vol. 1, Graylock Press, Rochester, NY, 1956.

[2] S. Berntein, *The theory of Probabilities*, Gastehizdat Publishing house, Moscow, 1946

[3] L. Birgé. *Approximation dans les espaces métriques et théorie d l'estimation.* Z. Wahrscheinlichkeitstheorie Verw. geb. **65**(1983), 181-237.

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

[5] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V.. Temlyakov, *Universal paper*, preprint

[6] A. Cohen, I. Daubechies, P. Vial, *Wavelets and fast wavelet transforms on an interval*, Appl. Comput. Harmon. Anal., **1** 1(1993), 54–81.

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

[8] A. Cohen, R. DeVore, R. Hochmuth, *Restricted nonlinear approximation*, Constr. Approx., **16** (2000), 85–113.

[9] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, *Tree approximation and encoding* , ACHA **11**(2001), 192-226.

[10] F. Cucker and S. Smale, *On the mathematical foundations of learning theory*, Bulletin . Amer. Math. Soc., **39** (2002), 1-49.

[11] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Regional Coference Series in Applied Mathematics, SIAM, Philadelphia, 1992.

[12] A. Dembo and O.Zeitouni, *Large deviation techniques and applications* Springer, (1998)

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

[14] R. DeVore, R. Howard, and C. Micchelli, *Optimal non-linear approximation*, Manuskripta Math. **63** (1989), 469-478.

[15] R. DeVore and R. Sharpley, *Besov spaces on domains in* $\mathbb{R}^d$, Trans, Amer. Math. Soc., **335** 1(1993), 843–864.

[16] R. DeVore and B. Lucier, *Wavelets*, Acta Numerica, **1** (1992), 1-56

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

[18] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage : Asymptopia ?* Journal of the Royal Statistical Society, Series B **57** (1995) 301-369.

[19] L. Györfi, M.Kohler, A. Krzyzak and H.Walk, *A Distribution -free Theory of Nonparametric Regression*, Springer Series in Statistics, 2002.

[20] A. Gushin *On Fano lemma and similar inequalities for the minimax risk.* To appear in Theor. Probability and Math. Statist.

[21] I. A. Ibragimov and R. Z. Hasminskii, Statistical estimation : asymptotic theory, Springer, New York, 1981.

[22] G. Kerkyacharian, and D. Picard, *Entropy, Universal coding, Approximation and Bases properties.* Constructive Approximation, **20** (2004), 1-37.

[23] B.S. Kashin and V.N. Temlyakov, *On a norm and approximation characteristics of classes of functions of several variables*, Metric theory of functions and related problems in analysis, Izd. Nauchno-Issled. Aktuarno-Finans. Tsentra (AFTs), Moscow, 1999, 69–99.

[24] S. Konyagin, V. Temlyakov, *Greedy approximation with regard to bases and general minimal systems*, Serdica Math. J., **28** (2002), 305-328.

[25] S. Konyagin, V. Temlyakov, *Some error estimates in learning theory*, IMI Preprint **05**(2004), 1-18

[26] S. Konyagin, V. Temlyakov, *The entropy in learning theory: error estimates*, IMI Preprint **09**(2004), 1-25.

[27] L. Le Cam, *Convergence of estimates under dimensionality restriction* Annals of Statistics, **1**(1973), 38-53.

[28] M. Ledoux and M. Talagrand *Probability in Banch spaces: Isoperimetry and Processes*. Sringer Verlag, New York, 1991.

[29] G. Lorentz, M. Von Golitschek, and Yu. Makovoz, *Constructive Approximation: Advanced problems*, Grundlehren vol. 304, Springer Verlag, Berlin, 1996.

[30] Y. Meyer, *Ondelettes et Operateurs I* Hermann, Paris (1990)

[31] T. Poggio and S. Smale *The mathematics of learning: dealing with data*, Notices of the AMS (to appear).

[32] S. Smale and D-X. Zhou *Estimating the approximation error in learning theory*, Analysis and Applications **1**(2003), 17–41.

[33] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, N.J., 1970.

[34] M. Talagrand, *New concentration inequalities in product spaces* , Invent. Math. **126**(1996), 505-563.

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

[36] V. Temlyakov, *Approximation by elements of a finite dimensional subspace of functions from various Sobolev or Nikol'skii spaces* , Mathematical Notes **43**(1988), 444–454.

[37] V. Temlyakov, *The best m-term approximation and greedy algorithms*, Adv. in Comput. Math., **8** 2(1998), 249–265.

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

[39] P. Wojtaszczyk, *Greedy algorithm for general biorthogonal systems*, J. Approx. Theory, **107** 2(2000), 293–314.

[40] P. Wojtaszczyk, *Projections and nonlinear approximation in the space $BV(I\!R^d)$*, Proc. London Math. Soc., **87** 3(2003), 471–497.

[41] Y. Yang and A, Barron *Information -Theoretic determination of minimax rates of convergence*, Annals of Statistics, **27**No. 5,(1999), 1564-1599.

[42] W.P Ziemer, *Weakly Differentiable Functions*, Springer–Verlag, New York, 1989.

Ronald A. DeVore, Dept. of Mathematics, University of South Carolina, Columbia, SC 29208, USA. email: devore@math.sc.edu
Gerard Kerkyacharian, Université Paris X-Nanterre, 200 Avenue de la République, F 92001 Nanterre cedex, France. email: kerk@math.jussieu.fr
Dominique Picard, Laboratoire de Probabilités et Modeles Aléatoires CNRS-UMR 7599, Université Paris VI et Université Paris VII, 16 rue de Clisson, F-750013 Paris, France. email: picard@math.jussieu.fr
Vladimir Temlyakov, Dept. of Mathematics, University of South Carolina, Columbia, SC 29208, USA. email: temlyak@math.sc.edu