Nonlinear Black-Box Models in System Identification : Mathematical Foundations

Anatoli Juditsky, Hakan Hjalmarsson,
Albert Benveniste, Bernard Delyon,
Lennart Ljung,
Jonas Sjöberg, and Qinghua Zhang *

June 27, 1995

Running title: Nonlinear Black-Box Models in Identification

*LL, HH, and JS are with Department of Electrical Engineering, Linköping University, S-581 83 Linköping, Sweden ; AB, BD, AJ, QZ are with IRISA/INRIA, Campus de Beaulieu, 35042 RENNES cedex, FRANCE ; e-mail: `name@isy.liu.se` and `name@irisa.fr`

Abstract

In this paper we discuss several aspects of the mathematical foundations of non-linear black-box identification problem. As we shall see that the quality of the identification procedure is always a result of a certain trade-off between the expressive power of the model we try to identify (the larger is the number of parameters used to describe the model, more flexible would be the approximation), and the stochastic error (which is proportional to the number of parameters). A consequence of this trade-off is a simple fact that good approximation technique can be a basis of good identification algorithm. From this point of view we consider different approximation methods, and pay special attention to *spatially adaptive* approximants. We introduce wavelet and "neuron" approximations and show that they are spatially adaptive. Then we apply the acquired approximation experience to estimation problems. Finally, we consider some implications of these theoretic developments for the practically implemented versions of the "spatially adaptive" algorithms.

Keywords: non-parametric identification, nonlinear systems, neural networks, wavelet estimators.

1 Introduction

The problem we are addressing in this paper is how to infer relationships between past input-output data and present/future outputs of a system when very little a priori knowledge is available. This is known as black-box modeling. There is a rich and well established theory for black-box modeling of linear systems, see *e.g.* [42] and [62]. It is not until the last few years that modeling and identification of non-linear systems have attracted a wide interest in the control community. Up to today, almost all attention has been concentrated on one single structure; neural networks. However, non-linear modeling have been studied for long in the statistics community where it is known under the label *non-parametric regression.* This area is quite rich and numerous methods exist. The purpose of this paper is to give an exposition of presently available techniques of non-linear modeling in a fairly unified and structured way. It is geared towards the mathematical foundations and exposes basic principles as well as presently available mathematical results. This exposition is, however, not exhaustive, neither with respect to presentation of existing methods nor with respect to mathematical results. In the companion paper [61] the user aspects and the algorithmic aspects are extensively discussed.

1.1 Basic mathematical problems

The basic problem that we will address throughout this paper is now precisely stated. We first state the general problem and then we discuss some features of dynamic system modeling.

The general problem

Problem 1 (non-parametric regression) *Let (X, Y) be a pair of random variables with values in $\mathcal{X} = R^d$ and $\mathcal{Y} = R$ respectively. A function $f : \mathcal{X} \mapsto \mathcal{Y}$ is said to be the* ***regression function of*** *Y* ***on*** *X if*

$$E(Y|X) = f(X) \ . \tag{1}$$

A typical case is $Y = f(X) + e$*, where* e *is zero mean and independent of* X*. For* $N \geq 1$*,* $\widehat{f}_N$ *shall denote an estimator of* f *based on the random sample* $\mathcal{O}_1^N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ *of size* N *from the distribution of* (X, Y)*, i.e., a map*

$$\widehat{f}_N \quad : \quad \mathcal{O}_1^N \mapsto \widehat{f}_N \left(\mathcal{O}_1^N, .\right) \tag{2}$$

where, for fixed $\mathcal{O}_1^N$*,* $x \mapsto \widehat{f}_N(\mathcal{O}_1^N, x)$ *is an estimate of the regression function* $f(x)$*. The family of estimators* $\widehat{f}_N$*,* $N \geq 1$ *is said to be* ***parametric*** *if* $\widehat{f}_N \in F$ *for all* $N \geq 1$*, where* F *is some set of functions which are defined in terms of a* fixed *number of unknown parameters. Otherwise the family of estimators* $\widehat{f}_N$*,* $N \geq 1$ *is said to be* ***non-parametric.***

For the sake of convenience, we shall often refer to X and Y as the *input* and *output* respectively (although they do not need to be such in actual applications).

Intuitively, the difference between the output and the regression function, $Y - E(Y|X)$, is the part of the output that cannot be predicted from past data.

Two typical problems are considered in the statistical literature, namely the

- *non-parametric regression with random design (or sampling),* where it is assumed that the variables X_i are random, independent, and identically distributed on $[0, 1]^d$ with density $g(x)$, and the
- *non-parametric regression with deterministic design (or sampling),* where it is assumed that the input variables X_i are nonrandom ; the simplest case of deterministic design is the *regular design,* where the inputs X_i form a regular grid (for instance, $f : R \to R$ and $X_i = i/N$).

In the remainder of this section we consider the random design only.

Non-parametric regression with dynamics.

Consider the following dynamical system :

$$Y_i = f(\Phi_i) + e_i, \qquad i = 1, \ldots, N \ ,$$

where $Y_i \in R$ and $\Phi_i \in R^d$ are observed, and e_i is a white noise as above. We assume that

$$\Phi_i = (Y_{i-1}, \ldots, Y_{i-m}; U_i, \ldots, U_{i-p}) \ , \tag{3}$$

where $U_i \in R$ denote the inputs ($m + p = d$). For example, if $\Phi_i = (Y_{i-1}, \ldots, Y_{i-d})$, then

$$Y_i = f(Y_{i-1}, \ldots, Y_{i-d}) + e_i \ . \tag{4}$$

In analogy with the corresponding parametric model we call this system a *non-parametric autoregression* or a *functional autoregression* of dimension d (NAR(d)). As an interesting application, we can consider a simple controlled NAR model for adaptive control :

$$Y_i = f(\Phi_i) + U_i + e_i \ , \tag{5}$$

where $\Phi_i = (Y_{i-1}, \ldots, Y_{i-m})$, and U_i is the control. The following question can be considered : How to choose the control (U_i) for the system (5) to track some reference trajectory $y = (y_i)$, or, at least, how to choose U_i in order to minimize EY_i^2, or, simply, to stabilize the system (5)? If the function $f(\Phi)$ was known, we could use the control

$$U_i = -f(\Phi_i)$$

to obtain $Y_i = e_i$. Clearly, this is a "minimum variance" control, since $EY_i^2 \geq \sigma_e^2 = Ee_i^2$. If f is unknown, a possible solution consists in performing non-parametric "certainty equivalence control": compute an estimate $\widehat{f}_N$ of the regression function f based on the observations of the input/output pair $(\Phi_i, Y_i - U_i)$, and then take

$$U_i = -\widehat{f}_i(\Phi_i) \ . \tag{6}$$

To analyze the certainty equivalence control (6), let us consider the control cost

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 = \frac{1}{N}\sum_{i=1}^{N}(f(\Phi_i) - \widehat{f}_i(\Phi_i))^2 + \frac{1}{N}\sum_{i=1}^{N} e_i^2 \ .$$

It is easily checked that

$$E(\widehat{f}_i(\Phi_i) - f(\Phi_i))^2 \to 0 \ \text{ when } i \to \infty \tag{7}$$

implies $EQ_N \to \sigma_e^2$, and $\widehat{f}_i(\Phi_i) - f(\Phi_i) \to 0$ a.s. implies $Q_N \to \sigma_e^2$ a.s. Thus condition (7) is instrumental in analyzing this problem, and we shall informally discuss how it can be guaranteed.

Denote by $\Phi_0^{i-1} = (\Phi_0, \ldots, \Phi_{i-1})^T$ the vector of all available inputs up to time $i-1$, and by $\varphi_0^{i-1} = (\varphi_0, \ldots \varphi_{i-1})^T$ the corresponding vector of integration variables. Let $P(\cdot)$ denote the distribution of the vector sequence (Φ_i) when driven by the unknown "true" model (5)–(6), let for some $1 \leq k \ll i$ $P_{\Phi_0^{i-k}}(\cdot)$ be a distribution of Φ_0^{i-k}, and let $p_{\Phi_i|\Phi_0^{i-k}}(\cdot)$ be a conditional density of the distribution of Φ_i given Φ_0^{i-k} (we assume that such a density exists). We have

$$E|\widehat{f}_i(\Phi_i) - f(\Phi_i)|^2 \sim \int |\widehat{f}_i(x) - f(x)|^2 \ p_{\Phi_i|\Phi_0^{i-k}}(x) \ dx \ P_{\Phi_0^{i-k}}(d\Phi_0^{i-k}) \ .$$

Note that, if the closed-loop system (5)–(6) is stable, one would reasonably take equal weights for the observations $\Phi_0, \ldots, \Phi_i$ in the estimate $\widehat{f}_i$. In such a case the estimate $\widehat{f}_i(\Phi)$ is asymptotically (as $i \to \infty$) slowly varying, i.e., $\widehat{f}_i \sim \widehat{f}_{i-k}$. On the other hand, the conditional density $p_{\Phi_i|\Phi_0^{i-k}}$ converges exponentially fast to the density $p_\Phi(x)$ of the invariant distribution of the Markov chain (Φ_i) (again, we suppose that the correspondent quantities exist). Thus we can write

$$E|\hat{f}_i(\Phi_i) - f(\Phi_i)|^2 \sim \int P_{\Phi_0^{i-k}}(d\Phi_0^{i-k}) \int |\hat{f}_{i-k}(x) - f(x)|^2 \ p_{\Phi_i|\Phi_0^{i-k}}(x) \ dx \ ,$$

and

$$E|\hat{f}_i(\Phi_i) - f(\Phi_i)|^2 \sim \int E \int |\hat{f}_i(x) - f(x)|^2 p_\Phi(x) dx \equiv R_i(\hat{f}, f).$$

Thus, as a conclusion, in any case, the crux in analyzing this adaptive minimum variance nonlinear control consists in getting bounds for the error in estimating the unknown function f. Note that the error measure we use in this case (the risk $R_i(\hat{f}, f)$) is rather specific: *the error norm is weighted with the density of observation.* Hence, in addition to proving consistency for the estimates, getting such bounds is an important question.

Remarks. The above discussion can be summarized as follows:

1. Non-parametric estimation of regression functions is instrumental in various problems such as adaptive identification and control.
2. Averaged L_p-norms of estimation error for various p's are natural candidates as a figure of merit.
3. Having bounds for the estimation error is of paramount importance. This has been illustrated on the adaptive control example.

1.2 Basic principles and limiting factors

There are two factors that limit the accuracy with which the regression function f can be determined. Firstly, only a finite number of observation points[1] $(x_k)_{k=1}^N$ are available. This means that $f(x)$, at other points x than those which are observed, must be obtained from the observed points by interpolation or extrapolation. Secondly, at the points of observation, $x_k, k = 1, ..., N$, $f(x_k)$ is observed with an additive noise $e_k = y_k - f(x_k)$. Clearly, the observation noises e_i introduce a random component in the estimation error. A general approach to the problem is the following: we first choose an approximation method, i.e. substitute the function in question by its approximation; then we estimate the parameters involved in this approximation. This way we reduce the problem of function estimation to that of parametric estimation, though the number of parameters we have to estimate is not bounded a priori and can be large. To limit the number of parameters some *smoothness* or *regularity* assumptions have to be stated concerning f. Generally speaking, smoothness conditions require that the unknown function f belongs to a particular restricted functional class.

To see how the stochastic and deterministic approximation errors are combined in the estimation problem, consider, for simplicity, the following example. Suppose that N noisy observations of an unknown function $f:\ R \to R$ are available:

$$Y_i = f(X_i) + e_i \ . \tag{8}$$

Suppose that f can be expressed in the form of some infinite expansion

$$f(x) = \sum_{j=0}^{\infty} \theta_j^* \ g_j(x) \ , \tag{9}$$

where $(g_j(x))$, $j = 0, ...$ is a known family of basis functions. This is our approximated model. Our "smoothness" assumption about f is that the coefficients θ_j^* are assumed to decrease in a certain way as $j \to \infty$. Thus the problem of estimating f reduces to the estimation of a suitable truncation of the vector of all parameters $\Theta^* = (\theta_0^*, \ldots,)^T$, using the observations (X_i, Y_i). An estimate $\widehat{\Theta}_N$ of Θ^* can be obtained by minimizing the following criterion

$$\sum_{k=1}^{N} \|Y_k - \widehat{\Theta}_N^T g(X_k)\|^2$$

[1]We distinguish between the random variables (Y_k, X_k) and the corresponding observations (y_k, x_k)

where $\|\cdot\|$ is some suitable norm (say Euclidean for instance), and the row vector $g(x)$ collects the components $g_j(x)$ corresponding to the θ_j selected in $\widehat{\Theta}_N$. Suppose that (X_k) is a realization of a stationary stochastic process, then the following holds asymptotically in $N \to \infty$:

$$E\|Y - \widehat{\Theta}_N^T g(X)\|^2 \quad \approx \quad \underbrace{E\|e\|^2}_{noise} + \underbrace{\|(\Theta - E\widehat{\Theta}_N)^T g(X)\|^2}_{bias} + \underbrace{E\|(E\widehat{\Theta}_N) - \widehat{\Theta}_N)^T g(X)\|^2}_{variance}$$

As we shall see later, usually, $E\widehat{\theta}_j = \theta_j^*$, and the bias term in the right-hand side does not depend on the data record. Thus the "bias" is in fact the approximation error due to the truncation of the infinite vector Θ^*. Let n be the dimension of $\widehat{\Theta}_N$. Increasing n would reduce the bias down to zero. But the variance term is typically $O(n/N)$, thus increasing n increases the variance term as well. The optimum occurs when both bias and variance terms are balanced. In Section 5 — Section 8 we shall see several examples of how this compromise is met. For a further empiric discussion of the bias/variance trade-off and its implications for identification, the reader is referred to [61].

The bottom line is that, to cope with the bias variance trade-off, it is important to use an efficient approximation technique, *i.e.* one that gives a small approximation error with few parameters. However, *which approximation method is effective depends on the particular function class to which the function is assumed to belong.* Since guessing appropriate function classes requires prior information which is hardly accessible to the engineer, *it is important to come up with approximation methods that are flexible and are as independent as possible from the particular function class.* This will be a recurring theme throughout the paper.

1.3 Structure of the paper

The paper is organized as follows. In the next section different smoothness classes are discussed together with associated approximation techniques from the point of view of their utility for estimation. "Spatially uniform" smoothness classes as well as classes of functions with sparse singularities are considered. For the latter classes it is shown that wavelets play an important role. There is a particular problem associated with the approximation of functions of a large number of input variables. How well observations data fill the input space decreases exponentially with the input dimension. Hence, it is necessary to further restrict the function class if approximations with reasonable accuracy is to be obtained for large input dimension and moderate sample sizes. This topic is given special attention in Section 3. Neural network, wavelets and other methods are discussed from this perspective.

The estimation problem is treated in Section 4 — Section 8. First, in Section 4, performance criteria are introduced ; note that new tools are required since concepts such as Cramer-Rao bounds and Fisher Information matrix are not appropriate for non-parametric estimation. Section 5 discusses the estimation of uniformly smooth functions. Classical techniques such as kernel, piecewise-polynomial and projection estimates are reviewed. Techniques to estimate non-uniformly smooth functions are dealt with in Section 6 while estimation of highly multivariate functions is considered in Section 8. Finally, as a conclusion, the gap between theory and everyday practice is discussed in Section 9.

2 Approximation in function spaces

As we have seen in Section 1, due to the bias/variance trade-off, the number of parameters used in the expression of the regression function for estimation has to be kept as small as possible. Thus approximation methods performing good approximation with few parameters will be preferred. Not surprisingly, the approximation method should be selected according to the prior assumptions on the function class.

¿From an application point of view, these classes can be categorized as being either classes of uniformly smooth functions or classes of locally spiky and jumpy functions. Many real-life nonlinear systems are smooth with sparse singularities, e.g., mechanical systems, chemical systems, etc. Thus classes of locally spiky and jumpy functions are important in practice. A basic problem is that one typically does not know which function class the unknown function belongs to. However, as we shall see, it is possible to come up with one single approximation scheme which has good approximation properties for a wide family of function classes covering both uniformly smooth classes and locally spiky and jumpy classes. In the sequel, we move progressively from simple to rather complex approximation problems.

2.1 Linear approximation schemes

We mean here approximation schemes that are linear in f, i.e., satisfy $(f+g)_n = f_n + g_n$ if f_n denotes the n-th order approximant of f. Typical examples are of the following form. We have some function space $\mathcal{F}$ and an increasing family of (closed) subspaces $\mathcal{F}_n$ converging to $\mathcal{F}$. Then we consider some norm $\|\cdot\|$ on $\mathcal{F}$. The nth-order *projection approximation* of $f \in \mathcal{F}$ is the $f_n \in \mathcal{F}_n$ minimizing the distance of f to the subspace $\mathcal{F}_n$. We call this type of approximations the projection approximants.

While such projections do not need to be associated with Hilbert space structures, the problem of determining an optimal approximation becomes particularly simple when the functional space is a Hilbert space. In that case, the best approximant is obtained by simple orthogonal projection of f onto some subspace. The following result can be obtained (cf, [53]).

Proposition 1 (Optimal approximants in Hilbert spaces) *Let $\{g_k\}_{k=1}^{\infty}$ be an orthonormal basis for a Hilbert space $\mathcal{H}$. Let $\{\mu_k\}_{k=1}^{\infty}$ be a sequence of non-increasing positive numbers. Consider the function class*

$$\mathcal{F} \quad = \quad \left\{ \sum_{k=1}^{\infty} c_k g_k \,:\, \sum_{k=1}^{\infty} \left| \frac{c_k}{\mu_k} \right|^2 \leq 1 \right\} . \tag{10}$$

Then for any $f \in \mathcal{H}$, and $f_n = \sum_{j=0}^{n} \mu_j g_j$, we have $\|f - f_n\|_{\mathcal{H}} \leq \mu_{n+1}$.

Comments:

1. The convergence rate in Proposition 1 in some sense cannot be improved: there are "bad" functions in the space $\mathcal{H}$ which lie at the distance μ_{n+1} from any n-dimensional linear subspace.

2. If $\mu_n \equiv 1, n = 1, 2, \ldots$, then $\mathcal{F} = \mathcal{H}$, and proposition 1 states that that the worst-case approximation error will not decrease when n is increased !! Indeed, in this case the corresponding "bad" function can be easily constructed: we can take simply $f = g_{n+1}$ — such a situation

occurs for instance if we take $s = 0$ in the example below. This cannot happen however when the coefficients μ_i vanish at a fixed rate when $i \to \infty$, i.e., when $\mathcal{F}$ is compact in $\mathcal{H}$.

3. The optimal approximation f_{n+1} can be computed recursively from f_n.

As an application of Proposition 1, consider the following class of functions. Let $\mathcal{W}_2^s(L)$ be the set of 1-periodic functions $f(x)$, which are defined by their Fourier series

$$f(x) = \sum_{j=1}^{\infty} c_j \varphi_j(x), \tag{11}$$

where $\varphi_{2k}(x) = \sqrt{2}\sin(2\pi kx)$ and $\varphi_{2k+1}(x) = \sqrt{2}\cos(2\pi kx)$, $k = 1, \ldots$ and where the Fourier coefficients satisfy

$$\sum_{j=1}^{\infty} |c_j|^2 (1 + |j|^{2s}) < L^2. \tag{12}$$

For this class of functions, the optimal approximant is given by $f_n = \sum_{j=0}^{n} c_j \varphi_j$ and the rate of convergence is n^{-s}. Notice that $\mathcal{W}_2^s(L)$ is a smoothness class. In fact (15) is one of the several equivalent definitions of the Sobolev class $\mathcal{W}_2^s(L)$, a particular case of the classes $\mathcal{W}_p^s(L)$, $p \geq 0$ which can be defined, for instance for $s < 1$, as the subset of the functions $f \in L_1$, such that

$$\frac{\|f(t+h) - f(t)\|_p}{|h|^s} \leq L\ . \tag{13}$$

As we have seen in Proposition 1, the best approximation of functions belonging to $\mathcal{W}_2^s$, when the error is measured in L_2-norm, is simply an orthogonal projection on some linear subspace. This is *not* true for other Sobolev classes. It can be shown that, for the class $\mathcal{W}_p^s$ with $p < 2$, the optimal projection on a subspace of dimension n converges to f with the rate $n^{-s'/d}$, where $s' = s - 1/p + 1/2$. We shall see, however, that there are different approximations, which exhibit a better convergence rate equal to $n^{-s/d}$. The difference between these two rates becomes significant for small p.

From the mathematical point of view the problem can be explained as follows: the Sobolev classes $\mathcal{W}_p^s(L)$ with $s - 1/p + 1/2 > 0$ are compact subsets of L_2. For $p \geq 2$ these subsets are convex. In contrast, when $p < 2$, these classes are not convex, and can hardly be approximated using linear (and thus convex) subspaces.

From the users point of view, the functions from the Sobolev classes $\mathcal{W}_p^s$ with small p are essentially classes of functions with sparse singularities, or classes with spatially "non-uniform" smoothness. Hence, if linear approximation schemes are used, the approximation rate for locally spiky and jumpy functions will be slower than for uniformly smooth functions.

Discussion. Roughly speaking, linear approximations schemes use subspaces (or basis functions) for approximation, which are independent from the particular function to be approximated. Thus the question arises if one could not do better by selecting the basis functions *adaptively,* i.e., depending on the function to be approximated. To illustrate this point consider the function $f(x) = 1_{\{0 \leq x < a\}}$ for some $0 < a < 1$. The Fourier coefficients of this function are

$$c_0 = a, \quad c_{2k} = \sqrt{2}\ \frac{\sin^2(\pi ka)}{\pi k}, \quad c_{2k+1} = \sqrt{2}\ \frac{\sin(\pi ka)\cos(\pi ka)}{\pi k},$$

From Proposition 1 we know that this function belongs to the subset $\mathcal{F}$ of $L^2[0,1]$ which consists of the functions such that the coefficients μ_k in decomposition (10) decrease slower than $k^{-1/2}$. This would provide a convergence rate not better than $n^{-1/2}$ for the orthogonal projection using n basis functions. However, one would naturally expect being able to design a procedure which focuses on detecting the edges of f, thus exhibiting a much better convergence rate. We shall see that such methods cannot be linear schemes but must be *spatially adaptive, i.e.* the basis functions must be able to adapt to the function to be approximated. Spatial adaptation is our next important topic. But first we introduce some suitable functional classes.

2.2 Besov spaces and classes of locally spiky and jumpy functions

A suitable family of spaces to deal with functions that are locally spiky and jumpy is that of the Besov spaces. This is *a family* of functional spaces indexed with 3 parameters (in a way that the family of Sobolev spaces W_p^s is indexed with two parameters: p and s). The interplay of these three indices gives to this family of spaces a great flexibility. For instance, for different combinations of the indices, we can obtain spaces of "uniformly" regular (or smooth) functions, as well as spaces of regular functions with sparse singularities. However, all Besov spaces possess the following important property (wavelets and wavelet expansions will be introduced later):

- ♡ norms in these spaces are easily evaluated using coefficients of the wavelet expansions of the functions of these spaces.

Let us move now on stating precise definitions. For the sake of clarity we consider only compactly supported [2] functions f: *supp* $f \subseteq [0,1]^d$, though all the definitions below can be generalized for non-compact and multi-dimensional cases (we recommend [65] and [66] as extremely complete presentations of the current state of the art in the theory of functional spaces).

For $f \in L_1$ and $M \in N$ we define the local oscillation of order M and radius t at the point $x \in [0,1]$ by

$$osc_M f(x,t) \;\; \triangleq \;\; \inf_P \frac{1}{t^d} \int_{|x-y|<t} |f(y) - P(y)| \, dy \; , \tag{14}$$

where the infimum is taken over all polynomials P of degree less than or equal to M. This quantity measures the quality of the local fit of f by polynomials on balls of radius t. Select $p, q > 0$, $s > d(p^{-1} - 1)$, and take $M = \lfloor s \rfloor$. The following set of functions :

$$\mathcal{B}^s_{pq} = \left\{ f \in L_{1 \wedge p} : \; \|f\|_{\mathcal{B}^s_{pq}} = \|f\|_p + \left(\sum_{j=1}^{\infty} (2^{js} \| \; osc_M f(x, 2^{-j}) \|_p)^q \right)^{1/q} < \infty \right\} \; , \tag{15}$$

(with the usual modification for p or $q = \infty$) is identical to the *Besov spaces* of functions [4], and it is shown in [65] that $\| \cdot \|_{\mathcal{B}^s_{pq}}$ is equivalent to the classical Besov norm.

Comments:

[2] *supp* f will be used to denote the support of f, *i.e.* the set where f is non-vanishing

1. The triple parameterization using s, p, and q provides a very accurate characterization of smoothness properties. As usual for Hölder or Sobolev spaces, the index s indicates how many derivatives are smooth. Then, for larger p, $\|f\|_{\mathcal{B}^s_{pq}}$ is more sensitive to details. Finally, index q has no useful practical interpretation, but it is a convenient instrument that serves to compare Besov spaces with the more usual Sobolev spaces $\mathcal{W}^s_p$, as indicated next. It is interesting to notice that the indicator functions of intervals belong to the spaces $\mathcal{B}^s_{s^{-1}\infty}$ for all $s > 0$, this illustrates our claim in the title of this subsection.

2. It can be shown that (cf [65]) for $s \geq 0$, $0 \leq p, q \leq \infty$:

   - The family of Besov spaces includes some more classical spaces. for s non integer, Hölder classes $\mathcal{C}^s = \mathcal{B}^s_{\infty\infty}$, and Sobolev spaces $\mathcal{W}^s_2 = \mathcal{B}^s_{22}$;
   - $\mathcal{B}^s_{pq} \subset \mathcal{B}^{s'}_{p'q'}$ if $p' \geq p$, $q' \geq q$, $s' \leq s - \frac{d}{p} + \frac{d}{p'}$ (strict inequality if $p = \infty$);
   - $\mathcal{B}^0_{pq} \subseteq L_p \subseteq \mathcal{B}^0_{pq'}$ where $q = 2 \wedge p$ and $q' = 2 \vee p$;
   - $\mathcal{B}^s_{pp} \subset \mathcal{W}^s_p \subset \mathcal{B}^s_{p2}$ for $p \leq 2$;
   - $\mathcal{B}^s_{p2} \subset \mathcal{W}^s_p \subset \mathcal{B}^s_{pp}$ for $p \geq 2$.

   In particular, if $s > d/p$, then $\mathcal{B}^s_{pq} \subset \mathcal{C}$.

2.2.1 Spline approximations in Besov spaces

We consider the d-dimensional case and $supp\ f \subseteq [0,1]^d$. *Free knots spline approximations* have been analyzed in [52] (Theorems 7.3 and 7.4) using Besov spaces. Recall that a function f_n is called a spline function on $[0,1]$ of order k with with n knots if $f_n \in \mathcal{C}^{k-2}$ and there exist points (knots) $0 = x_0 < x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n = 1$ such that f_n is an algebraic polynomial of degree $k-1$ in each interval (x_{i-1}, x_i). Therefore, a spline is a smooth piecewise polynomial function. Free knot spline approximants are not linear schemes. The following result shows that any function from Besov space can be nicely approximated by splines. More surprisingly, any function which have a good spline approximant belongs to a certain Besov space.

We first state the so-called *Jackson inequality* for spline approximations. Consider $f \in \mathcal{B}^s_{pq}, p, q > 0$. Then there exists a spline function with n free knots f_n such that the following bound holds:

$$\|f_n - f\|_u \leq C(s,p,q)\ n^{-s}\ \|f\|_{\mathcal{B}^s_{p\infty}} , \tag{16}$$

where u satisfies $s - 1/p + 1/u > 0$. The converse bound is provided by the *Bernstein inequality:* For any $f \in L_u$, $s - 1/p + 1/u = 0$, $u < \infty$,

$$\|f\|_{\mathcal{B}^s_{pp}} \leq C(s,p,q) \left(1\ +\ n^s\ \inf_{f_n} \|f - f_n\|_u\right) ,$$

where the infimum ranges over the set of spline functions f_n of order $k \geq s+2$ with n free knots. A similar result holds in the multi-dimensional case and for nth-order *rational fraction approximations,* see Theorems 7.3 and 8.3 in [52].

In contrast, approximations using *fixed linear subspaces* perform poorly in Besov spaces. Consider some increasing family $(\mathcal{L}_n)$ of n-dimensional linear subspaces of $L_u, u > p$. Let f_n denote the linear projection of $f \in \mathcal{B}^s_{pq}$ on $\mathcal{L}_n$ using the L_u-norm. Then for any such family $(\mathcal{L}_n)$, there exists a least favorable f such that the following lower bound holds :

$$\|f - f_n\|_u \geq C \; n^{-s'} \; \|f\|_{\mathcal{B}^{s'}_{uu}} \; , \tag{17}$$

where $s' = s - 1/p + 1/u$. Note that the Sobolev injections stated at the end of section 2.2 imply that the same bounds hold for Sobolev classes $\mathcal{W}^s_p$.

Consider again the example of the indicator function $f(x) = 1_{\{0 \leq x < a\}}$. It can be easily verified that $f \in \mathcal{B}^s_{s^{-1}\infty}$ for any $s > 0$. On the one hand, it follows from (16) that f can be approximated using splines with n knots with asymptotic L_2-error of order $o(n^{-l})$ for any $l < \infty$ as $n \to \infty$. On the other hand, by (17), linear approximations of the same function have an L_2-error of order $O(n^{-1/2})$, where n is the dimension of the linear subspace. This remark would make rational approximations or splines with free knots very attractive for approximation in Besov spaces. Unfortunately, in the above result only the existence of such approximations is stated, and they are very hard to compute, for example, the optimal positioning of the knots of the spline approximation is very hard to find. It is amazing that *wavelet approximations are as good as spline ones, but are much more easily constructed.* We discuss this next.

2.2.2 Wavelet approximations in Besov spaces

The original objective of the the theory of wavelets is to construct orthogonal bases of $L_2(R)$ of the form $(\psi(2^j x - k))_{j,k}$, i.e. the bases which are constituted by translations and dilations of the same function ψ. It is preferable to take ψ localized and regular. We refer the reader to [61] for the attractive computational features of orthonormal wavelet bases, and we concentrate here on the properties that are useful to understand how they perform for function approximation.

The principle of wavelet construction is the following: first construct a function $\varphi \in L_2(R)$ such that

(S1) the functions $\varphi(x - k)$ mutually orthogonal for k ranging over Z.

(S2) φ is a *scale* function, i.e. there is a sequence $h_k \in l_2$ such that

$$\varphi(x) = \sqrt{2} \sum_{k \in Z} h_k \varphi(2x - k) \; . \tag{18}$$

This is an important step, several constructions of φ have been proposed (cf [9]). Next we define the *wavelet*

$$\psi(x) = \sqrt{2} \sum (-1)^k \; h_k \; \varphi(2x - k) \; . \tag{19}$$

It can be shown that the family $\{\psi(2^j x - k) \, , \, j \in N, k \in Z\}$ constitutes an orthonormal basis of $L_2(R)$. The crux in proving this property is to verify that, for any j_0, the family $\{\varphi(2^{j_0} x - k), \psi(2^j x - k), k \in Z, j \geq j_0\}$ also forms an orthogonal basis of $L_2(R)$; this is achieved mainly by using the algebraic properties **(S1)**, **(S2)**, (19). If φ and ψ are compactly supported, they give us a local description, at different scales j, of the considered function:

$$f(x) = \sum_{k \in Z} \langle f, \varphi_{j_0 k} \rangle \varphi_{j_0 k}(x) + \sum_{j \geq j_0, k \in Z} \langle f, \psi_{jk} \rangle \psi_{jk}(x)$$

where $\varphi_{jk}(x) = \varphi(2^j x - k)$, and $\langle .,. \rangle$ denotes inner product in L^2.

In what follows we assume that φ is a compactly supported piecewise continuous scale function satisfying the following condition :

$$\exists r > 0 \quad : \quad \varphi \in \mathcal{B}^r_{u\infty} \ , \tag{20}$$

and we move to the multidimensional case. Starting from φ one can construct the corresponding orthonormal basis of $L_2(R^d)$, i.e. the functions $\Phi, \Psi^{(i)},\ i = 1, ..., 2^d - 1$ such that for any $f \in L_2(R^d)$ we have the formal expansion

$$\begin{aligned} f &= \sum_k \alpha_{0k}\ \Phi_{0k} + \sum_{j=0}^{\infty} \sum_{k \in Z^d} \sum_{l=1}^{2^d-1} \beta^{(l)}_{jk}\ \Psi^{(l)}_{jk} \\ \alpha_{jk} &= \langle f, \Phi_{jk} \rangle, \quad \beta^{(l)}_{jk} = \langle f, \Psi^{(l)}_{jk} \rangle \ . \end{aligned} \tag{21}$$

Here

$$\left\{ \Phi_{0k},\ \Psi^{(l)}_{jk} \right\}\ ,\ 0 \leq j < \infty, k \in Z^d, 1 \leq l \leq 2^d - 1$$

is the corresponding orthonormal basis, formed by dilations and multi-dimensional translations of Φ and $\Psi^{(l)}$. For details of definitions, the reader is again referred to [61].

Wavelet approximations. We first state a result [45] [35] concerning functions that satisfy Hölder type conditions. Recall, that a function f is called *Hölder continuous with exponent s at point x_0*, written $f \in \mathcal{C}^s_{x_0}$, if there is a polynomial P of degree at most $\lfloor s \rfloor$ such that [3]

$$|f(x) - P(x - x_0)| \leq C|x - x_0|^s \ .$$

If f is Hölder continuous, with exponent $s < r$ (r is the regularity of φ at x_0, see (20)), then there exists $C < \infty$ such that, for any integer $j > 0$,

$$\max_{\{k\ :\ x_0 \in\ supp\ \Psi_{jk}\}} \langle f, \Psi_{jk} \rangle \ \leq \ C\ 2^{-j(s+d/2)} \ . \tag{22}$$

Conversely, if (22) holds and f is known to be $\mathcal{C}^\varepsilon_{x_0}$ for some $\varepsilon > 0$, then

$$|f(x) - P(x - x_0)| \leq C|x - x_0|^s \ \log \frac{2}{|x - x_0|} \ .$$

This result states that local smoothness of Hölder type can be characterized with the vanishing rate of the wavelet coefficients in the neighborhood of this point. This property is specific to the wavelet transform, and does not hold for other orthogonal bases. As we will see a similar property holds in Besov spaces. We have the following result (c.f. Theorem 4 in [60]):

Theorem 1 (Besov norms and wavelet decompositions) *Let $r > s > d(1/u - 1)$ and φ be a scale function satisfying conditions (20). For any $f \in \mathcal{B}^s_{pq}$ define*

$$\|f\|_{spq} = \left(\sum_k |\alpha_k|^p \right)^{1/p} + \left(\sum_{j=0}^{\infty} \left[2^{j(s+d/2-d/p)} \|\beta_{j\cdot}\|_p \right]^q \right)^{1/q} \tag{23}$$

[3] recall that $\lfloor s \rfloor$ denotes the largest integer $\leq s$.

and $\|\beta_{j\cdot}\|_p = (\sum_{l,k} |\beta_{jk}^{(l)}|^p)^{1/p}$, *see (21) for the definition of coefficients* $\alpha_k = \alpha_{0k}$ *and* $\beta_{jk}^{(l)}$. *Then (23) is equivalent to the norm of Besov space* $\mathcal{B}_{pq}^s$, *i.e., there exist constants* C_1 *and* C_2, *independent of* f, *such that*

$$C_1 \; \|f\|_{\mathcal{B}_{pq}^s} \;\leq\; \|f\|_{spq} \;\leq\; C_2 \; \|f\|_{\mathcal{B}_{pq}^s} \; . \tag{24}$$

Theorem 1 states that norms in Besov spaces are suitably evaluated using orthonormal wavelet decompositions. This fact can be used to obtain very efficient approximations.

We now indicate how such a wavelet approximation of f can be constructed. Consider the full wavelet decomposition of f:

$$f(x) \;=\; \sum_{k\in Z} \alpha_{0k} \; \Phi_{0k}(x) + \sum_{j=0}^{\infty} \sum_{k\in Z^d} \sum_{l=1}^{2^d-1} \beta_{jk}^{(l)} \; \Psi_{jk}^{(l)}(x) \; . \tag{25}$$

1. Keep the projection of f on the subspace V_0, this corresponds to the left most sum in (25). When f and Φ are both compactly supported this requires computing only a fixed amount of coefficients, say m. And then
2. Select in the second (triple) sum those coefficients β_λ, $\lambda = (i, j, k)$ with largest absolute value, denote by Λ the set of the $n - m$ so selected wavelet coefficients. Finally
3. Add $n - m$ detail terms $\beta_\lambda \Psi_\lambda$ to the sum taken in step 1.

This procedure yields the approximation

$$w_n(x) \;=\; \underbrace{\sum_{k} \alpha_{0k} \; \Phi_{0k}(x)}_{\substack{m \text{ coeffs.} \neq 0 \\ (f,\Phi \text{ compact. supp.})}} \; + \underbrace{\sum_{j=0}^{\infty} \sum_{k\in Z^d} \sum_{l=1}^{2^d-1} \beta_{jk}^{(l)} \; \Psi_{jk}^{(l)}(x)}_{keep\ the\ largest\ n-m\ coeffs.} \tag{26}$$

and the following theorem provides corresponding approximation bounds.

Theorem 2 (DeVore, Jawerth, Popov, [13]) *Consider* $f \in \mathcal{B}_{pp}^s$, $s, p > 0$ *and* $s - d/p + d/u \geq 0$. *Let* w_n *denote the approximation (26) of* f. *If the scale function satisfies condition (20), then*

$$\|f - w_n\|_u \;\leq\; C(s,p) \; n^{-s/d} \; \|f\|_{\mathcal{B}_{pp}^s} \tag{27}$$

holds. If, in addition, u *satisfies* $s - d/p + d/u = 0$, $u < \infty$, *and it is a priori known that* $f \in L_u$, *then the following converse bound holds.*

$$\|f\|_{\mathcal{B}_{pp}^s} \;\leq\; C(s,p,q) \; \left(1 \;+\; n^{s/d} \; \|f - w_n\|_u\right) \; .$$

Comments: This result is quite remarkable for the following reasons:

1. This approximation procedure gives the same rate of approximation for a wide variety of different Besov spaces (those satisfying $s - d/p + d/u \geq 0$). Especially, it is not necessary to *a priori* know the extent of localized singularities, *i.e.*, index p.

2. When certain norms are used to measure the approximation error, and for functions with localized singularities, the approximation error tends to zero much faster if (26) is used than if linear approximation is performed.This follows by comparing (27) with the rate $n^{-s+d/p-d/u}$ which is generic for linear approximations for the cases where $0 < p \leq u$ and $p \leq u$.

Also, in the wavelet decomposition of a function with sparse singularities (e.g., $f \in \mathcal{B}^s_{pq}$, $p < 2$), only a small number of basis functions are important, and the other ones can be neglected. In contrast to spaces of uniformly smooth functions, *it is not necessarily the first basis functions that should be used.* Let us go once more back to our example $f(x) = 1_{\{0 \leq x < a\}}$. Consider the wavelet decomposition of this function using a compactly supported wavelet $\psi(x)$ such that $\int \psi(x)dx = 0$. It is evident that the coefficient β_{jk} vanishes for any wavelet $\psi_{jk}(x)$ which does not cross the (local) singularities of f. Thus if we consider the projection of f on the subspace V_j, only $O(j)$ coefficients of the decomposition significantly differ from zero (among 2^j candidates).

3 Approximation in high-dimensional spaces

3.1 The curse of dimensionality

The accuracy of an approximation depends on how densely observation points fill the input space. Thus having enough data points for good estimation would require the sample size to grow exponentially with the input dimension. This is referred to as the *curse of dimensionality*, a phrase coined by Bellman [3]. The curse of dimensionality is exhibited explicitly in the results (16) and (27) where the rate of convergence has order $n^{-s/d}$. For large input dimension d, this is exceedingly slow.

3.2 Function classes of lower effective dimension

There are basically two ways to deal with the curse of dimensionality. Either to accept that a huge amount of data is necessary or to restrict the function class further. The latter means that, instead of having the dimension visible in the convergence rate, it will be hidden behind the function class. Kolmogorov [38] proved that every continuous function on $[0, 1]^d$ can be represented as the additive superposition of continuous one-dimensional functions. Lorentz [43] gave an explicit scheme: Every continuous function f on $[0, 1]^d$ can be written as

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{2d+1} g_j \left(\sum_{k=1}^{d} h_{jk}(x_k) \right) \tag{28}$$

for some continuous univariate functions (g_j). Moreover, the functions (h_{jk}) can be taken to be universal, *i.e.* they do not depend on f. Unfortunately, these results are not of great help for approximation, since the above functions g_j are usually extremely irregular even for a smooth f function [4].

However, one way to bound the function class in its "effective dimension" is suggested by the generic decomposition (28). Introduce the variables $z_j = \sum_{k=1}^{d} h_{jk}(x_k)$ which, since the h_{jk} are known, can be precomputed. The function f can then be written as

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{2d+1} g_j(z_j) \tag{29}$$

and the problem is to approximate $2d + 1$ univariate functions g_j. Using m basis functions to approximate each g_j gives an approximation error of the order m^{-s} under usual smoothness assumptions on g (cf Section 2). Taking $m = n/(2d+1)$, the total number of basis functions is n and the total approximation error will be of the order $(2d + 1)(n/(2d + 1))^{-s}$ which, for large n, is of order n^{-s} and is much better than the above quoted $n^{-s/d}$.

Projection onto one-dimensional subspaces is the crux of the *Projection Pursuit Algorithm*, developed in [30] (a very good review of these results can be found in [32]), which consider estimates of f in the form

$$\widehat{f}_N(x) = \sum_{j=1}^{M} \widehat{g}_j(\alpha_j^T x) \ ,$$

where (α_j) are unit vectors and each $\alpha^T x$ may be thought of as a projection of x. The j-th term $\widehat{g}_j(\cdot)$ is constant along $\alpha_j^T x = c$ and so is often called a *ridge function*: the estimate at a given point can be thought of as based on the averages over certain (in general adaptively chosen) strips $\{x : |\alpha_j^T x - t_i| \leq \varepsilon\}$. Other examples are *Recursive Partitioning* [46], [6] and related methods (c.f., for instance [29] with discussion). These methods are derived from some mixture of statistic and heuristic arguments and give impressive results in simulations. Their drawback lies in the almost total absence of any theoretical results on their convergence rates. We refer the reader to the above references for additional information.

3.3 Neural networks

For a review of neural networks see the companion paper [61]. The following result was recently published in [2], it is the most accurate theoretical result on the neural network-based approximations today. Let $\sigma(x)$ be a sigmoidal function (i.e. a bounded measurable function on the real line for which $\sigma(x) \to 1$ as $x \to \infty$ and $\sigma(x) \to 0$ as $x \to -\infty$). Consider a compactly supported function f with *supp* $(f) \subseteq [0,1]^d$, and assume that

$$C_f = \int_{R^d} |\omega| \ |\widehat{f}(\omega)| \ d\omega < \infty \ , \tag{30}$$

where $\widehat{f}(\omega)$ denotes the Fourier transform of f. The main result of [2] can be roughly stated as follows: there exists an approximation f_n of the compactly supported function f, of the form

$$f_n(x) = \sum_{i=1}^{n} c_i \sigma(a_i^T x + t_i) + c_0 \tag{31}$$

[4] This was already noted in the original paper by Kolmogorov.

(note that f_n is *not* compactly supported), such that

$$\|(f_n - f)\ 1_{[0,1]^d}\|_2 \ \leq\ 2\sqrt{d}\ C_f\ n^{-1/2}\ . \tag{32}$$

This result provides an upper bound of the minimum distance (in L_2-norm) between any f satisfying condition (30) and the class of all neural networks of size not larger than n. In the same article, the upper bound (32) is compared with the best achievable convergence rate for any linear approximation in class (30). It is shown that a lower rate for linear projections is $n^{-1/d}$, compare this with the much better rate $n^{-1/2}$ for neural networks, especially for large dimension d.

Comment It is not easy to relate the function class defined by the condition (30) and more usual smoothness classes. For instance, one can show that if $f \in W^{(d+1)/2+\epsilon}(L)$ for some $\epsilon > 0$ and $L < \infty$, then (30) holds. Note that that the same rate was obtained for "linear" projection estimators in section 2.1. On the other hand, it can be shown that even in the Sobolev class $W^{(d+1)/2}$ there are "bad" functions, which does not verify (30).

This gives us an idea that the neural approximator outperform usual linear estimators on some special rather restricted functional classes. *What are these classes?* In the article [2] some classes are listed. In particular, if $f(x) = g(x^T a)$ for some $a \in R^d$, $|a| = 1$, i.e. f(x) is a ridge function, then the Fourier transform $\widehat{f}(\omega)$ is concentrated in the direction of a and conditions (30) implies that $\int |\omega| \widehat{g}(\omega) d\omega < \infty$.

It is much easier, however, to answer the question what are the classes on which neural nets behaves badly. Consider a class of function which are spherically symmetric, i.e. $f(x) = g(|x|)$ for some $g:\ R \to R$. Using spheric coordinates ρ, ν (here $\rho = |x|$ and ν is a vector on the unit sphere and ρ is the radius), we get from (30)

$$\int_{R^d} |\omega|\ |\widehat{f}(\omega)|\ d\omega \quad = \quad \int_0^\infty \rho |\widehat{g}(\rho)| d\rho |S_\rho| \tag{33}$$

where $|S_\rho|$ is the surface of the d-dimensional sphere of radius ρ, which is

$$S_\rho = \frac{\rho^{d-1}\pi^{d/2} d}{\Gamma(d/2+1)}$$

, where $\Gamma(\cdot)$ is the standard Γ-function. We conclude from (30) and (33) that $g(\cdot)$ should verify

$$\int \omega^d |\widehat{g}(\omega)| d\omega < \infty.$$

This is a hard assumption, and it implies that the d-th derivative of g is bounded. We know (cf. section 2.2) that for such functions the rate n^{-1} can be attained by spline or wavelet approximations (and many other classical methods). The rate $n^{-1/2}$ which is stated in (32) for neuron approximations is really not good in this case. These two examples illustrates the the following simple idea: the neural nets are not always good approximants. They behave badly on certain functional classes and outperform local estimators in some particular situations. This duality between local and "semi-local" methods has been discussed and developed in [17].

When coming back to Barron's result, it should be noted that no result is available which takes advantage of possible improved smoothness of the unknown function f. An iterative algorithm for the construction of the approximation (31) is also proposed. The true problem of system identification, i.e., that of neural network training based on noisy input/output data, is not addressed in this paper.

3.4 Wavelets

Note that in the orthonormal wavelet expansion

$$f(x) = \sum_k \alpha_{0k} \Phi_{0k}(x) + \sum_{ljk} \beta_{jk}^{(l)} \Psi_{jk}^{(l)}(x) ,$$

the dilation and translation parameters – 2^{-dj} and k do not depend on the function to expand, only the linear weights α_{jk} and $\beta_{jk}^{(l)}$ depend on f. Suppose that we are able to construct an "adaptive wavelet basis", i.e., with dilations and translations depending on the function f. The wavelet expansion of f using these basis functions is expected to use less wavelets, and thus we expect it to be more convenient for estimation purposes. To obtain such a basis we can *discretize the continuous wavelet transform* (36) which is given by the following theorem.

Theorem 3 *Let ψ and φ be radial* [5] *functions satisfying*

$$\forall \omega \in R^d \ : \ \int_0^\infty a^{-1} \widehat{\varphi}(a\omega) \widehat{\psi}(a\omega) da = 1 \tag{34}$$

where we recall that $\widehat{\varphi}(\omega)$ denotes the Fourier transform of function $\varphi(x)$. Then for any function $f \in L_2(R^d)$, the following formulae define an isometry between $L_2(R^d)$ and a subspace of $L_2(R^d \times R_+)$ [9]:

$$u(a,t) = a^{d-1/2} \int f(x) \ \varphi(a(x-t)) \ dx \tag{35}$$

$$f(x) = \int u(a,t) \ \psi(a(x-t)) \ a^{d-1/2} \ da \ dt \ . \tag{36}$$

Here $a \in R^+$ and $t \in R^d$ are respectively the dilation and translation factors.

We present the following algorithm proposed in [12]. Consider the continuous wavelet transform (36), which we rewrite as

$$\begin{aligned} f(x) &= \int u(a,t) \ \psi(a(x-t)) \ a^{d-1/2} \ da \ dt \\ &= \int \psi(a(x-t)) \ \mathrm{sign}(u(a,t)) \ a^{(d-1)/2} \ |u(a,t)| \ da \ dt \\ &= \frac{1}{C} \int \psi(a(x-t)) \ \mathrm{sign}(u(a,t)) \ w(a,t) \ da \ dt \end{aligned}$$

where we have renormalized $u(a,t)$ by a constant factor C so that the function $w(a,t) = C a^{(d-1)/2} \, |u(a,t)|$ can be considered as a probability density. Then we draw n independent random samples $(a_i, t_i)_{i=1,\ldots,n}$ from a distribution with density $w(a,t)$. Then we build

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n a_i^{d/2} \ \psi(a_i(x-t_i)) \ \mathrm{sign}(u(a_i,t_i)) \ , \tag{37}$$

which, thanks to the law of large numbers, converges to the true wavelet transform. Some faster implementations of this algorithm are given in [12]. Improving this estimate by some "bootstrapping" like technique, yields the following approximation result.

[5] A function φ is radial if $\varphi(x)$ depends only on $|x|$; this implies that $\widehat{\varphi}(\omega)$ also depends only on $|\omega|$.

Theorem 4 ([12]) *ψ is any radial wavelet function such that there exists a related radial function φ which satisfies condition (34). Let p, μ, l, s be real numbers satisfying*

$$1 < p < \left(1 - \frac{s-l}{d}\right)^{-1}, \qquad \mu = \min\left(1 - \frac{1}{p}\ ,\ \frac{1}{2}\right)$$

and f be a function of the Sobolev space $W_1^s(R^d)$; then, for any $n > 0$ there exists a function f_n of the form

$$f_n(x) = \sum_{i=1}^{n} u_i\ \psi(a_i(x - t_i)) \tag{38}$$

such that

$$\|f_n - f\|_{W_p^l} \leq C n^{-\mu} \|f\|_{W_1^s}\ .$$

In particular, if $s > d/2$ then

$$\|f_n - f\|_2 \leq n^{-1/2} C \|f\|_1^s\ . \tag{39}$$

Comment: Theorem 4 provides us with an upper bound for the rate of approximation when adaptive dilation/translation sampling is used to discretize the continuous wavelet transform. We should compare this rate with rates of convergence for approximations based on *fixed* dilation/translation sampling. For those approximations the best rate which can be attained for $p = 2$ and $s - d/2 = \varepsilon > 0$, would $n^{-\varepsilon/d}$, which is much slower than the rate (39).

Note also that in this case the rate of convergence for the "shrunk" wavelet approximants is also $n^{-1/2-\varepsilon}$. Thus these two types of nonlinear approximations have almost the same rate of convergence.

3.5 Breiman's Hinging Hyperplanes

We now briefly discuss a recent technique due to Leo Breiman [5], which practically combines some advantages of neural networks (in particular the ability to handle very large dimensional inputs) and of constructive wavelet based estimators (availability of fast training algorithms). Breiman's technique is an elegant way of identifying piecewise linear models based on data collected from an unknown nonlinear system, see [63] for the use of such models in control. Following [5], we call *hinge function* a function $y = h(x)$, $x \in R^d$ which consists of two hyperplanes continuously joined together, i.e., an open book. If the two hyperplanes are given as

$$y = \langle \beta^+, x \rangle + \beta_0^+, \quad y = \langle \beta^-, x \rangle + \beta_0^-,$$

where $\langle ., . \rangle$ denotes scalar product in Euclidean spaces, then an explicit form for the hinge function is either

$$\begin{aligned} h(x) &= \max(\ \langle \beta^+, x \rangle + \beta_0^+\ ,\ \langle \beta^-, x \rangle + \beta_0^-\ )\ , \\ \text{or} \qquad h(x) &= \min(\ \langle \beta^+, x \rangle + \beta_0^+\ ,\ \langle \beta^-, x \rangle + \beta_0^-\ )\ . \end{aligned}$$

It is proved in [5], using the methods by Barron [2] that there is a constant C such that for any n there are hinge functions $h_1, \ldots, h_n$ such that

$$\|f - \sum_{i=1}^{n} h_i 1_{[0,1]^d}\|_2 \leq C n^{-1/2} \tag{40}$$

for any f such that

$$\int_{R^d} |\omega|^2 |\widehat{f}(\omega)| d\omega < \infty \ , \tag{41}$$

i.e., Breiman's hinge model is as efficient as neural networks for the L_2-norm. Notice that, as in the neural network case, the condition (41) limits the function class by reducing the effective dimension, *c.f.* the comment after (32). However, as indicated at the beginning of this section, no convergence rate is given for models identified from noisy data (the bound (40) is not a convergence rate for identification, but only a rate of approximation of a given function by some finitely parameterized class of approximants). Further details are given in the companion paper [61] concerning effective procedures for hinging hyperplanes estimation, we shall not discuss them any further.

4 Performance measures for non-parametric estimators

With the approximation results from the previous sections at hand, we are now ready to move on to analyze the behavior of estimation algorithms from noisy data. The performance analysis of non-parametric estimation algorithms and/or identification procedures is much more difficult than for parametric estimation. The following specific issues are important:

1. What plays the role of Cramer-Rao bound and Fisher Information Matrix in our case? Recall that the Cramer-Rao bound reveals the best performance one can expect in identifying the unknown parameter θ from sample data arising from some parameterized distribution $p_\theta, \theta \in \Theta$, where Θ is the domain over which the unknown parameter θ ranges. In the non-parametric case, lower bounds for the best achievable performance are provided by *minimax risk functions.* We shall introduce these lower bounds and discuss associated notions of optimality.

2. For lower bounds, what is the class of systems on which best achievable performance is considered, is another important issue. For non-parametric representations of linear systems, $L_2, L_\infty, H_2, H_\infty$, with their associated norm are typical spaces to work with. For (even static) nonlinear systems, however, the choice is much wider. How wide should be the class $\mathcal{F}$ of systems in consideration, what kind of smoothness should be required? Are we interested in the behavior of the estimate at one particular point x of interest, or are we interested in the global behavior of the estimate? Different distance measures should be used in these two different cases.

4.1 Lower bounds for best achievable performance.

In order to compare different non-parametric estimators it is necessary to introduce suitable figures of merit. It seems first reasonable to build on the mean square deviation (or mean absolute deviation) of some semi-norm [6] of the error, we denote it by $\|\widehat{f}_N - f\|$. The following semi-norms are commonly used in non-parametric regression : $\|f\| = \left(\int f^p(x)dx\right)^{1/p}$, $0 < p < \infty$ (L_p-norm), $\|f\| = \sup_x |f(x)|$ (uniform norm, $\mathcal{C}$- or L_∞-norm), $\|f\| = |f(x_0)|$ (absolute value at a fixed point x_0). Then we consider the *risk function*

$$R_{a_N}(\widehat{f}_N, f) = E\left[a_N^{-1}\|\widehat{f}_N - f\|\right]^2 , \tag{42}$$

where a_N is a normalizing positive sequence. Letting a_N decrease as fast as possible so that the risk still remains bounded yields a notion of a convergence rate. Let $\mathcal{F}$ be a set of functions which contains the "true" regression function f, then the maximal risk $r_{a_N}(\widehat{f}_N)$ of estimator $\widehat{f}_N$ on $\mathcal{F}$ is defined as follows :

$$r_{a_N}(\widehat{f}_N) = \sup_{f \in \mathcal{F}} R_{a_N}(\widehat{f}_N, f) \ .$$

If the maximal risk is used as a figure of merit, the optimal estimator $\widehat{f}_N^*$ is the one for which the maximal risk is minimized, i.e., such that [7]

$$r_{a_N}(\widehat{f}_N^*) = \min_{\widehat{f}_N} \sup_{f \in \mathcal{F}} R_{a_N}(\widehat{f}_N, f) \ .$$

We call $\widehat{f}_N^*$ *the minimax estimator* and the value

$$\min_{\widehat{f}_N} \sup_{f \in \mathcal{F}} R_{a_N}(\widehat{f}_N, f)$$

the minimax risk on $\mathcal{F}$. Notice that this concept is consistent with the mini-max concept used in the definition of n-widths in approximation theory in Section 2.

The construction of minimax non-parametric regression estimators for different sets $\mathcal{F}$ is a hard problem. Presently, it has only solved asymptotically (for large samples) for some special cases (see, for instance, [25], [26], [27]). However, letting a_N decrease as fast as possible so that the minimax risk still remains bounded yields a notion of a best achievable convergence rate, similar to that of parametric estimation. More precisely, we state the following definition :

Definition 1 (lower rate and minimax rate of convergence)

1. *The positive sequence a_N is a* ***lower rate of convergence for the set $\mathcal{F}$ in the semi-norm*** $\|\cdot\|$ *if*

$$\liminf_{N\to\infty} r_{a_N}(\widehat{f}_N^*) = \liminf_{N\to\infty} \inf_{\widehat{f}_N} \sup_{f \in \mathcal{F}} E\left[a_N^{-1}\|\widehat{f}_N - f\|\right] \ \geq \ C_0 \tag{43}$$

for some positive C_0.

This notion can be refined as follows.

[6] a semi-norm is a norm, except it does not satisfy the condition : $\|f\| = 0$ implies $f = 0$.

[7] to properly understand the statement to follow, the reader should pay attention to definition (2) of an estimator.

2. *The positive sequence a_N is called* ***minimax rate of convergence*** *for the set $\mathcal{F}$ in semi-norm $\|\cdot\|$, if it is a lower rate of convergence, and if, in addition, there exists an estimator $\widehat{f}_N^*$ achieving this rate, i.e., such that*

$$\limsup_{N\to\infty} r_{a_N}(\widehat{f}_N^*) \;<\; \infty \;.$$

The inequality (43) is a kind of negative statement that says that no estimator of function f can converge to f faster than a_N. Thus, a coarser, but easier approach consists in assessing the estimators by their convergence rates. In this setting, by definition, optimal estimators reach the lower bound as defined in (43) (recall that the minimax rate is not unique: it is defined to within a constant).

4.2 Some negative results.

¿From the discussion in Section 1, it should be evident that it is the assumed smoothness class that dictates the minimax rate of convergence. Generally it holds that the larger the class of functions, the slower the convergence rate. Devroye and Györfi [15] [14], have proven the following result[8]. Consider the following classes of functions on R:

$\mathcal{F}^*$: the class of all functions f such that $f(x) = 0$ for $x > 1$ or $x < 0$, and $|f(x)| \le C$ for $x \in [0,1]$.

$\mathcal{F}_0^*$: the class of all continuous functions $f \in \mathcal{F}^*$.

$\mathcal{F}_\infty^*$: the class of all functions $f \in \mathcal{F}^*$ having all continuous derivatives on $[0,1)$ (notice that the interval is right open).

Let $\widehat{f}_N$ be an arbitrary estimate of f. Then for the classes $\mathcal{F}^*$, $\mathcal{F}_0^*$ and $\mathcal{F}_\infty^*$ defined above (we denote them generically by $\mathcal{F}$),

$$\sup_{\mathcal{F}} \limsup_{N\to\infty} E\left[a_N^{-1}\int_0^1 |\widehat{f}_N(x) - f(x)|dx\right] = \infty$$

for any positive sequence $a_N \to 0$.

Thus, *no convergence rate does exist for any of the above classes $\mathcal{F}^*$, $\mathcal{F}_0^*$ and $\mathcal{F}_\infty^*$.* In other words, the convergence can be arbitrary slow, depending on the unknown function or density f to be estimated! It is a natural consequence of the fact that the above classes $\mathcal{F}^*$, $\mathcal{F}_0^*$ and $\mathcal{F}_\infty^*$ are too rich: they contain functions which are extremely difficult to approximate. *In other words, in order to obtain any interesting rate of convergence, smoothness conditions should be imposed.*

4.3 Some positive results.

Let us now concentrate on the case of deterministic uniform design, i.e., the input data X are uniformly sampled in the considered interval. The following result in the case of regular design can be acknowledged to [34] (for the random design case, see [64], [39]).

[8] Note, however, that convergence can sometimes be proved without any smoothness assumption [16].

Theorem 5 *Let us consider the Hölder class $\mathcal{C}^s(L)$ on $[0,1]^d$, closely related to the Sobolev classes. is the The Hölder class $\mathcal{C}^s(L)$ is the the family of functions $f(x)$, $x \in [0,1]^d$ defined by* [9]

$$\mathcal{C}^s(L) = \left\{ f : \ |f^{(k)}(x) - f^{(k)}(x')| \leq L|x-x'|^{s-k}, \text{ for any } x, x' \in [0,1]^d \right\} \quad , k = \lfloor s \rfloor \ . \tag{44}$$

Consider

$$\|g\| = \left(\int |g(x)|^p dx \right)^{1/p}, \quad 0 < p < \infty \qquad or \quad \|g\| = |g(x_0)| \ .$$

Then $N^{-\frac{s}{2s+d}}$ is a lower rate of convergence for the class $\mathcal{C}^s(L)$ in the semi-norm $\|\cdot\|$. Furthermore, $\frac{N}{\ln N}^{-\frac{s}{2s+d}}$ is a lower rate of convergence for the class $\mathcal{C}^s(L)$ in the norm $\|g\| = \sup_{x \in [0,1]} |g(x)|$.

Note that to obtain the correct rate of convergence for the distance at a fixed point x_0, the corresponding Lipschitz property is required at x_0 only. Similar results hold when the class $\mathcal{C}^s(L)$ is replaced by the class $\mathcal{W}_p^s(L)$, $p \geq 2$ (see (13)). Then $N^{-\frac{s}{2s+d}}$ is also a lower rate of convergence for this class in the L_p-norm of the error.

5 Estimation in classes of uniformly smooth functions

Throughout this section, Problem 1 is considered. The discussion in Section 1.2 and Section 2 give the required background for understanding how to perform estimation of the unknown f function using a fixed basis function expansion. Let us take the Sobolev class $\mathcal{W}_2^s(L)$ and our model (8), (9). In this case we take as (g_j) the Fourier basis, and to obtain the optimal approximation with n basis functions, we can simply take first n terms of the expansion (9) or (11). This gives a certain maximal bias error, *cf.* the first term in (10). In this particular example, this error is of order $n^{-s/d}$. The parameters in the function expansion (9) are estimated via empirical means based on N noisy observations. The mean square error of the estimate of each coefficient is $O(1/N)$. Thus the total mean square error of the estimate will be, as usual, the sum of the stochastic part and of the bias due to the approximation error : this yields $O(n/N) + O(n^{-2s/d})$. The optimal choice for n balances these two terms : $n = N^{\frac{1}{2s+d}}$. This choice for n yields a quadratic error of order $N^{-\frac{2s}{2s+d}}$. This is the typical scheme that is followed even in the cases where the basis function expansion is not as explicit as in this example.

The estimators we consider in this section are *linear*, i.e. they can be written

$$\widehat{f}_N(x) = \sum_{i=1}^{N} Y_i \ W_{N,i}(x) \ , \quad W_{N,i}(x) = W_{N,i}(x, X_1, \ldots, X_N) \tag{45}$$

where we recall that $\mathcal{O}_1^N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ is the given random sample observation. It is only the weights $W_{N,i}(x)$ that may differ from estimator to estimator. This means that such an estimator satisfy $\widehat{f+g} = \widehat{f} + \widehat{g}$; notice, however, that functions f, g, and their estimates, are generally nonlinear as functions of their input x. Linear estimators build the folklore of non-parametric estimation : kernel estimators, projections on linear subspaces of functions, are typical instances we shall describe. We shall then discuss, both practically and theoretically, some severe practical limitations of linear estimators. Roughly speaking, linear estimators are suitable for systems with "uniform smoothness" ; systems with sparse singularities (e.g., hard limiters, quantizers, some mechanical systems) are poorly handled. This follows from the discussion at the end of Section 2.1.

[9] $\lfloor s \rfloor$ denotes the maximal integer $k < s$.

5.1 Kernel estimators for regression functions and densities

Kernel estimators were first proposed by Nadaraya and Watson in 1964 [48] and [69]. The Nadaraya-Watson kernel estimator is an interpolation procedure. It is given by

$$\widehat{f}_N(x) = \frac{\sum_{i=1}^N Y_i \; K\left(\frac{X_i - x}{h_N}\right)}{\sum_{i=1}^N K\left(\frac{X_i - x}{h_N}\right)} \; , \tag{46}$$

where (h_N) is a sequence of positive numbers, $h_N \to 0$ as $N \to \infty$, and K is a function on R satisfying

$$\lim_{|u|\to\infty} |u||K(u)| = 0, \quad \int_{-\infty}^{\infty} |K(u)| du < \infty, \quad \sup_{u\in R} |K(u)| < \infty, \quad \int_{-\infty}^{\infty} K(u) = 1 \; . \tag{47}$$

The positive number h_N is called the *bandwidth* and the function K satisfying (47) is called a *kernel*; in fact, h_N is better interpreted as a scaling factor. Clearly, the Nadaraya-Watson estimator is linear, and has the form (45). Typical examples of kernels are $K(u) = (1/2)\, 1_{\{|u|\leq 1\}}$ (rectangular window kernel), and $K(u) = (1/\sqrt{2\pi}) \exp(-|u|^2/2)$ (Gaussian kernel), etc... Usually K is chosen to be an even function.

The idea of kernel estimation is simple, let us explain it for the case of the rectangular kernel in one dimension. In this case the estimator (46) is a simple moving average with equal weights: the estimate at point x is the average of observations Y_i corresponding to X_i's belonging to the "window" $[x - h_N, x + h_N]$. If $h_N \to \infty$ then the estimator tends to $N^{-1} \sum_i Y_i$, the average of all observations, and thus for functions f which are far from being constant, the bias becomes large. If h_N is very small (say, smaller than the pairwise distance between sample points X_i) then the estimator reproduces the observations: $\widehat{f}_N = Y_i$. In this extremal case the variance of the error becomes high. Thus increasing h_N tends to increase the bias of estimator, while reducing h_N leads to a larger variance. The optimal choice for h_N corresponds to an equal balance between bias and variance.

Also closely related to estimator (46) is the Parzen-Rosenblatt kernel estimator for densities. Let $X_1, \ldots, X_N$ be independent and identically distributed random variables with common density $f(x), x \in R^d$. The Parzen-Rosenblatt estimator of density $f(x)$ is a suitably smoothed histogram. It is defined as [51], [58]

$$\widehat{f}_N(x) = \frac{1}{N h_N^d} \sum_{i=1}^N K\left(\frac{X_i - x}{h_N}\right) \; , \tag{48}$$

where d is the state-space dimension of X and K is a kernel as in (47). Kernel estimate (46) can be easily derived from the Parzen-Rosenblatt one. Recall definition (1) of the regression function, take the Parzen-Rosenblatt estimator (48) for the joint density $f(x, y)$ of (X, Y) and denote it by $\widehat{f}_N(x, y)$. Then, replacing, in formula

$$f(x) = \frac{\int y \; f(x, y) \; dy}{\int f(x, y) \; dy} \; ,$$

$f(x)$ and $f(x, y)$ by their corresponding Parzen-Rosenblatt estimates, yields Kernel estimate (46).

We now state a sample of results about the properties of kernel estimates for the d-dimensional case when it is known a priori that f belongs to the Hölder class $\mathcal{C}^s(L)$ (see (44) for the definition of $\mathcal{C}^s(L)$).

We acknowledge Rosenblatt [59] for the first two statements of it, though it probably belongs to the earlier folklore of non-parametric statistics.

Theorem 6 ([59]) *Let $\widehat{f}_N$ be a kernel estimate with bandwidth h_N such that $h_N \to 0$ and $Nh_N \to \infty$, with kernel K satisfying $\int x^j K(x)dx = 0 \quad for \quad j = 1, \ldots, k$. Here, x^j denotes any product of the form $x_1^{j_1} x_2^{j_2} \ldots x_d^{j_d}$ where $j_1 + j_2 + \ldots + j_d = j$, and $x_1, \ldots, x_d$ are the coordinates of x. Assume that the observations X_i are independent and identically distributed on $[0,1]^d$ with density $g(x) \geq c > 0$, $g \in \mathcal{C}^s(L)$, and that the noise satisfies $Ee_i = 0$ and $Ee_i^2 \leq \sigma_e^2 < \infty$. Then*

1. *Uniformly over $f \in \mathcal{C}^s(L)$ and $x \in [0,1]^d$, we have the pointwise bound*

$$E|\widehat{f}_N(x) - f(x)|^2 \leq C \left(L^2 h_N^{2s} + \frac{\sigma_e^2}{N h_N^d} \right) . \tag{49}$$

The optimal value of h_N which minimizes the right-hand side of (49) is given by

$$h_N = \left(\frac{\sigma_e^2}{L^2 N} \right)^{1/(2s+d)} . \tag{50}$$

For this value of h_N

$$E|\widehat{f}_N(x) - f(x)|^2 \leq C L^{2/(d+2s)} \left(\frac{\sigma_e^2}{N} \right)^{2s/(2s+d)} .$$

2. *If we consider instead the global error measure $E\|\widehat{f}_N - f\|_2^2$, using again the same optimal value (50) for h_N yields the same bound, uniformly over $f \in \mathcal{C}^s(L)$.*

Comments:

1. As expected from the above informal discussion concerning the rectangular kernel, the bound for the estimation error variance given on the right hand side of (49) is decomposed into *bias* + *variance* terms. And, as expected, the optimal choice of h_N in (50) exactly balances these two terms.

2. Note that we have both pointwise and global bounds, which reflects the local nature of kernel estimates.

3. The properties of the Parzen-Rosenblatt algorithm of density estimation are identical when the unknown density f satisfies $f \in \mathcal{C}^s(L)$. Note that, since *supp* $f \subseteq [0,1]^d$, the L_1-norm of the error (restricted to the $[0,1]^d$) is dominated by the L_2-norm. So we get from the second statement of the theorem

$$E\|\widehat{f}_N - f\|_1^2 \leq C L^{2/(d+2s)} \left(\frac{\sigma_e^2}{N} \right)^{2s/(2s+1)}$$

provided h_N is chosen as in (50).

4. Often the following recursive version of the kernel estimator is considered, [24], [50]:

$$\widehat{f}_n(x) = \Gamma_n^{-1}(x)\left(\sum_{i=0}^{n} Y_i h_i^{-d} K\left(\frac{X_i - x}{h_i}\right)\right) \quad \text{if } \Gamma_n(x) \neq 0 \text{ and } \widehat{f}_n(x) = 0 \text{ if } \Gamma_n(x) = 0 \ ,$$
$$\Gamma_n(x) = \sum_{i=0}^{n} h_i^{-d} K\left(\frac{X_i - x}{h_i}\right) \ ,$$

or

$$\begin{aligned}\widehat{f}_n(x) &= \widehat{f}_{n-1}(x) + \Gamma_n^{-1}(x)\left(Y_n - h_n^{-d} K\left(\frac{X_n - x}{h_n}\right)\widehat{f}_{n-1}\right) \ , \\ \Gamma_n(x) &= \Gamma_{n-1}(x) + h_n^{-d} K\left(\frac{X_n - x}{h_n}\right) \ . \end{aligned} \tag{51}$$

In this form the algorithm resembles very much the recursive Least Squares algorithm for estimating the parameters of linear models. When the bandwidth is such that $h_i = h\ i^{-\alpha}$ for some $0 < \alpha < 1$, the properties of the algorithm (51) in the static regression problem are essentially the same as those of the "off-line version" (46). In [50], [56] and [24] this algorithm was used to identify stable non-parametric autoregression models of the form (4), and the convergence of this estimator was proved. Furthermore, the same algorithm was used to provide the estimates of $\widehat{f}_n$ in the closed loop system (5)–(6), and the stability of such an adaptive control scheme was proved — [50] and [56] consider essentially one-dimensional case, and in [24] the general multi-dimensional case is studied.

5.2 Piecewise-polynomial estimators

Another non-parametric regression estimator which is commonly used is the piecewise-polynomial one. The idea is the same as for the kernel estimator, though the averaging is made over *bins* (i.e., small cubes) of fixed size δ_N rather than in h_N-neighborhood of the current point x. It is also closely related to the *radial-basis function (RBF) networks* with rigid location for the radial functions, see [54], [68]. The simplest example of this method is the piecewise-constant estimator or *regressogram*. The value of the estimate $\widehat{f}_N$ in each bin equals the average of observations Y_i such that corresponding X_i belong to the bin. For the sake of clarity we consider the one-dimensional case.

The piecewise-polynomial estimator can be formally defined in terms of the following optimization problem. Let $\delta_N \to 0$ be a positive sequence, and we assume that $\delta_N^{-1} = M$ is an integer. Define $u_l = l\delta_N$, $l = 0, \ldots, M$, and divide the interval $[0,1]$ into M cubes (bins) of the form $U_1 = [0, u_1)$, $U_2 = [u_1, u_2)$, $\ldots, U_M = [u_{M-1}, 1]$, so each bin has length δ_N. Set $F(x) = (1, x, \frac{x^2}{2}, \ldots, \frac{x^k}{k!})^T$ and, for each bin $U_l, l = 1, \ldots, M$, solve for $\theta \in R^{k+1}$ in the least squares sense the system of equations

$$Y_i = \theta^T F\left(\frac{X_i - u_{l-1}}{\delta_N}\right), \ X_i \in U_l \tag{52}$$

and denote by $\widehat{\theta}_{N,l}$ the corresponding solution. Then the piecewise-polynomial estimate $\widehat{f}_N$ of order k in each bin U_l is expressed as

$$\widehat{f}_N(x) = \widehat{\theta}_{N,l}^T\ F\left(\frac{x - u_{l-1}}{\delta_N}\right), \ \ x \in U_l \tag{53}$$

The value δ_N is called the *binwidth.* As for the bandwidth h_N of kernel estimate, the binwidth tunes the smoothness: larger δ_N leads to a higher bias, and smaller δ_N results in a higher variance. In order for the least-squares problem in (53) to be non-degenerate we require that the number of points X_i in each bin is larger than $k+1$.

Stone, [64] has proved a result similar to Theorem 6 for this type of estimate (see (44) for the definition of the Hölder space $\mathcal{C}^s(L)$). We state this result in the general d-dimensional case. Assume that the observations X_i satisfy the assumptions of Theorem 6. Let $\widehat{f}_N$ be a piecewise polynomial estimate of order $k = \lfloor s \rfloor$, with binwidth $\delta_N \to 0$ and $N\delta_N \to \infty$ as $N \to \infty$. Then *statement 1 of Theorem 6 holds with binwidth δ_N substituted for the bandwidth h_N.*

Comments:

1. Note that, unlike for Kernel estimates, piecewise polynomial estimates compute projections on the fixed set of functions $F\left(\frac{x-u_{l-1}}{\delta_N}\right)$, $x \in U_l$ (the l-th bin). The same remark holds for the projection estimate to follow.

2. As can be seen, piecewise polynomial and kernel estimates have the same asymptotic accuracy when $N \to \infty$.

3. If f is a smooth function (i.e., $s \geq 1$), the optimal number of bins is $n_\delta \sim \delta_M^{-1}$ and is much less than the number of observations ($n_\delta \sim N^{1/3}$ for $s = 1$). This number is equivalent to the memory size required to implement the algorithm: to reconstruct the estimate, $k = \lfloor s \rfloor$ coefficients are necessary. Thus, if N is large, this algorithm offers significant advantage, in terms of memory requirements, over kernel estimates in which all measurements should be kept to reconstruct $f(x)$. Also, computing (52)–(53) is of lower computational burden than computing (46). These two points make the piecewise polynomial estimate more attractive.

4. Unfortunately there is no reasonable recursive version of the estimate $\widehat{f}_n$. Although on can use the recursive least squares algorithm to compute linear regression coefficients $\widehat{\theta}_{N,l}$ in (53), the derivations quickly become messy, because the number M of bins depends on N, and so does the number of equations in the algorithm.

5.3 Projection estimates

Another class of function estimates was introduced by Cencov [7], who called them *projection estimates.* The idea consists in expanding the unknown function into its "empirical" Fourier series. Consider the Sobolev class $\mathcal{W}_2^s(L)$ of functions $f(x)$ (11) but now defined for $x \in [0,1]^d$. In this case (11) becomes

$$f(x) = \sum_{|j|=1}^{\infty} c_j \varphi_j(x), \tag{54}$$

where $j = (j_1, \ldots, j_d)$ is a multi-index, $x = (x^1, \ldots, x^d)^T$, $\varphi_j(x) = \varphi_{j_1}(x^1) \times \ldots \times \varphi_{j_1}(x^d)$, $\varphi_1 \equiv 1$, $\varphi_{2k}(x) = \sqrt{2}\sin(2\pi k x)$ and $\varphi_{2k+1}(x) = \sqrt{2}\cos(2\pi k x)$, $k = 1, \ldots$. The condition (12) remains the same

$$\sum_{j=1}^{\infty} |c_j|^2 (1 + |j|^{2s}) < L^2. \tag{55}$$

We assume that the *input X is uniformly distributed.* We construct the estimate $\widehat{f}_N$ as follows:

$$\widehat{f}_N(x) = \sum_{j=1}^{m} \widehat{c}_j^N \varphi_j(x), \tag{56}$$

where m is the "model order", and the empirical estimates $\widehat{c}_j^N$ of Fourier coefficients

$$\widehat{c}_j^N = \frac{1}{N}\sum_{i=1}^{N} Y_i \varphi_j(X_i) \tag{57}$$

are substituted for the true ones c_j, $j = 1, \ldots, m$. Note that the assumption that X is uniformly distributed has been used. Note also that the estimate (56)–(57) is linear (cf. (45)) with weights given by

$$W_{N,i}(x) = \sum_{j=1}^{m} \frac{1}{N}\ \varphi_j(x)\ \varphi_j(X_i).$$

Cencov, [7] has proved the following counterpart of statement 1 of Theorem 6 : Let $\widehat{f}_N$ be a projection estimate. Then, uniformly over $f \in \mathcal{W}_2^s(L)$ and $x \in [0,1]^d$,

$$E\|\widehat{f}_N(x) - f(x)\|_2^2 \leq C\left(L^2 m^{-2s} + \frac{\sigma_e^2 m^d}{N}\right). \tag{58}$$

The optimal order m of the model is

$$m = \lfloor \left(\frac{L^2 N}{\sigma_e^2}\right)^{1/(2s+d)} \rfloor\ , \tag{59}$$

it balances bias and variance error estimates, and yields the bound

$$E\|\widehat{f}_N(x) - f(x)\|_2^2 \leq C L^{2/(d+2s)}\left(\frac{\sigma_e^2}{N}\right)^{2s/(2s+d)}. \tag{60}$$

The following result, due to Ibragimov and Khas'minskij [34], provides a global uniform bound. Take

$$m = \lfloor (\frac{N}{\ln N})^{1/(2s+d)} \rfloor$$

for the model order (note that this is slightly different from (59)). Then, uniformly over $f \in \mathcal{C}^s(L)$ (the class $\mathcal{C}^s(L)$ is defined in (44)), it holds that

$$E\|\widehat{f}_N - f\|_\infty^2 \leq O\left(\frac{\ln N}{N}\right)^{2s/(2s+d)}. \tag{61}$$

Comments:

1. Projection estimates have the same rate of convergence (up to a constant) as kernel or piecewise polynomial ones.

2. The bound (58) for the quadratic error of the algorithms appears rather naturally if we consider the following argument: when we approximate $f \in \mathcal{W}_2^s$ using m terms of its Fourier decomposition, the approximation error is $O(m^{-2s/d})$. Furthermore, the stochastic error in each term is of order $O(N^{-1})$. This simple calculus can be repeated for any non-parametric estimate. Obviously, it is beyond our possibilities to reduce the stochastic component of the error. On the other hand, the bias part depends on the method we choose to approximate the function (piecewise polynomial, trigonometric series, etc.) with, and this choice of approximant is of primary importance.

3. From the computational point of view, projection estimates are more attractive than piecewise polynomial estimates, since it uses an orthonormal basis of functions (the Fourier basis), which dramatically simplifies the computation of the least-squares estimates $\widehat{c}_j$ of the Fourier coefficients c_j, cf. (57) with (52).

5.4 Selecting model complexity

As we have seen, the convergence of the estimates strongly depends on the choice of the bandwidth h_N for kernel estimator, the model order m for the projection estimator, and the binwidth δ_N (or, equivalently, the "model order" $M = \delta^{-1}$) for piecewise polynomial estimator. *These design parameters depend on the parameters of the smoothness class $\mathcal{C}^s(L)$ or $\mathcal{W}_2^s(L)$, which are a priori unknown* — see definition (44) of this class and the use of parameters (s, L) in Theorem 6 and corresponding results for others estimators. Even if some information about the smoothness parameter s is available, the knowledge of the value L is of importance when the data sample is of bounded length. Let us illustrate this with the following example, where the input x is scalar. Consider the problem of estimating a function $f(x)$ in additive white noise e, with $\sigma_e^2 = 1$. Assume that f has support $[0, 1]$, that all its derivatives are continuous, and that $f(1/2) = 1$, $f(0) = f(1) = 0$. Note that in this case, typically, $\sup_x |f^{(s)}(x)| \approx s^s$, i.e., higher order derivatives become very large in uniform bound. In this case the bounds in Theorem 6 are of order $a_N(s) = (s/N)^{2s/(2s+1)}$ when the parameter is selected for the smoothness s. Assume that the size of the observation sample is $N = 10000$, then $a_N(2) = 0.0110$, $a_N(3) = 0.0095$, but we have already $a_N(4) = 0.0122$ (the value of s which minimizes a_N is $s \approx 3.4814$ with $a_N(s) \approx 0.00946$). This illustrates the fact that the tightest bound is not obtained by taking the largest possible s, but rather by selecting the most favorable pair (s, L), which obviously is much more difficult.

Given that we only have in practice samples of finite size N, we shall not try to estimate the most favorable pair (s, L), but we shall proceed differently. The model order (or bandwith, or binwidth, depending on the different estimates) shall be estimated from data using a procedure usually referred to as the *generalized cross validation* (GCV) test. GCV procedures were studied for kernel (see, for instance, [57], [31]), spline ([40], [8]), and projection estimates (c.f. [55], [41]). Let us consider, for instance, the procedure for the projection estimates [10]. To make the model order explicit in formula (56) we shall write $\widehat{f}_{m,N}$ instead of $\widehat{f}_N$. Set $S_{m,N}^2 = N^{-1} \sum_{i=1}^N \|Y_i - \widehat{f}_{m,N}(X_i)\|^2$. As

[10] In fact, a similar result holds for the spline or piecewise polynomial ones.

for the prediction error variance estimate in parametric prediction error methods, $S^2_{m,N}$ is a *biased* estimate of the error. Thus one cannot minimize $S^2_{m,N}$ with respect to m directly : the result of such a brute force procedure would give a function $\widehat{f}_{m_N,N}(x)$ which perfectly fits the noisy data, this is known as "overfitting" in the neural network literature. The solution rather consists in introducing a penalty which is proportional to the model order m, i.e., we search for m_N such that

$$m_N = \arg\min_{m\leq N}\left(S^2_{m,N} + \frac{2\sigma_e^2 m}{N}\right). \tag{62}$$

This technique is clearly equivalent to the celebrated Mallows-Akaike criterion [44], [1]. The following result, due to Polyak and Tsybakov [55], shows the consistency of this procedure. Assume that the Fourier coefficients of f in expansion (11) satisfy $|c_j| \leq \varepsilon_j$, $\sum_{j=1}^{\infty}\varepsilon_j < \infty$, $(j\,\varepsilon_j)$ is non-increasing, and σ_e^2 is known. Set $V_{m,N} = \|\widehat{f}_{m,N} - f\|_2^2$. Then for the estimate (56), (57), and (62), it holds that

$$\frac{V_{m_N,N}}{\min_m V_{m,N}} \to 1 \ \text{ a.e. as } N \to \infty \ .$$

Another "classical" adaptation approach is closely related to the problem of filtering of a Gaussian stationary process. The technique developed in [25], [27] for the projection estimates is often referred to as Efromovich-Pinsker filter. To understand the idea of the method let us consider the estimates $\widehat{c}_j$, $j = 1, ..., N$ of the Fourier coefficients c_j of f as in Section 5.3. If the observation noise (e_i) is independent Gaussian, then the errors $\widehat{c}_j^N - c_j$ are Gaussian and un-correlated (and thus independent), and the Wiener filter can be applied to the sequence $\widehat{c}_j$ to obtain the estimates $\tilde{c}_j$ of c_j:

$$\tilde{c}_j = \frac{c_j^2}{c_j^2 + \sigma_e^2/N}\widehat{c}_j. \tag{63}$$

It can be shown that this choice of Fourier coefficients yields the least possible asymptotic L_2 error among all "projection" estimates. Naturally, the exact values of the coefficients c_j are not available. To construct an adaptive filter Efromovich and Pinsker proposed to use instead of the filter coefficients their estimates, which are obtained in the following way: we divide the set of indices $j = 1, ..., N$ into m groups: $T_1 = \{j = 1, j = 2\}$, ..., $T_{k+1} = \{j = 2^k + 1, ..., j = 2^{k+1}\}$, ..., $T_m = \{j = 2^{m-1} + 1, ..., j = 2^m\}$ (we have supposed for the sake of clarity that $N = 2^m$). We set

$$\Lambda_k = \frac{\Theta_k}{\Theta_k + \sigma_e^2/N},$$

where

$$\Theta_k = |T_k|^{-1}\sum_{j\in T_k}(\widehat{c}_j^2 - \sigma_e^2/N)$$

(here $|T_k|$ is a cardinality of T_k). We put, finally,

$$\tilde{c}_j = \Lambda_k \widehat{c}_j$$

for all $j \in T_k$. It was shown (cf. [27]) that this adaptive filter is asymptotically equivalent to the optimal one (63).

Remark Note that the adaptive algorithms, though we have developed them starting with linear methods, are not linear anymore. Quite naturally, when trying to infer some additional information from the data we loose the linearity of the estimates.

6 Estimation in classes of locally spiky and jumpy functions

6.1 Spatial adaptivity

The estimation of functions with sparse singularities should naturally be based on function approximation in the corresponding smoothness classes. This was discussed in Section 2.2 — Section 3.5. During the last fifteen years this topic has been a very active field in the statistical community and has been characterized by successful practical applications but, oddly enough, an almost complete lack of theoretical results. Spatially adaptive methods include all sorts of neural networks, projection pursuit [30], classification and regression trees (CART) [6], Multivariate adaptive regression splines (MARS) [29], Variable Bandwidth Kernel methods [47], and others.

These methods implicitly or explicitly attempt to adapt the fitting method to the form of the function being estimated, by ideas like recursive dyadic partitioning of the space on which the function is defined (CART and MARS) and adaptively estimating a local bandwidth function (Variable Kernel Methods).

We discuss now some issues related to the problem of spatial adaptivity. In fact, so-called "spatially adaptive" methods address two different problems:

Estimating functions which have sparse singularities and otherwise are smooth. [11] An interesting approach consists in *finding a parameterized family of functional classes which*

1. *fits our prior knowledge about the smoothness of the function to be estimated, (in particular, that f is smooth everywhere, except at a sparse set of points), and*
2. *has associated with it an estimation technique which is minimax within these classes.*

It was the merit of David Donoho and Iain Johnstone [18] to recognize that Besov spaces, which play a central role in Yves Meyer's mathematical theory of wavelets [45], provide an adequate answer. They are perfectly suited to nonlinear systems which have sparse singularities and otherwise are smooth. The methods used can be qualified as local function expansions, they provide a combination of local averaging and short-range interpolation.

Handling geometric issues in the multi-dimensional case. If, for instance, the function of two variables $f(x, y)$ is "regular" in x (in extreme case $f(x, y) = g(y)$ does not depend on x) and "irregular" in y, then an "intelligent" estimation algorithm would approximate thoroughly f only in y-direction. This way the problem can be reduced to that of function estimation in 1-dimensional, with the correspondent improvement in the rate of convergence. This effect is often called "the dimensionality reduction" in the statistical literature. Starting from early 1980's a variety of techniques have been proposed in the statistics literature, which

[11] The CART and MARS algorithms, and the variable bandwidth kernel method were designed to handle this problem in multi-dimensional case. Surprisingly enough, the A.I. literature has proposed independently and at the same time different techniques with the same feature of "spatial adaptivity". These include various forms of neural networks [33].

exhibit this desirable feature of "dimensionality reduction". The already mentioned *Projection Pursuit Algorithm* and neural network algorithms are examples of this idea. This contrasts with wavelet and other local averaging procedures, in which the smoothing is done over small balls $\{x : |x - x_0| \leq \varepsilon'\}$. It was shown in [17] that, in a certain setting, projection based and local-averaging function estimates have complementary properties.

We will defer the discussion of the second issue to Section 8.

In the literature on non-parametric regression, the focus, so far, has been on locally adaptive bandwidths for kernel methods, see [67] for an example. Adaptive local linear regression is treated in [28].

6.2 Wavelet shrinkage algorithms

Wavelet shrinkage algorithms offer the dual advantage of being 1/ spatially adaptive (and thus practically efficient) and of low comparatively computational cost, and 2/ supported by a complete mathematical analysis. This dual feature is rather unique, so we now concentrate on this technique.

Non-parametric regression. Assume that N samples of input/output observations of the following system are available :

$$Y_i = f(X_i) + w_i \ ,$$

where (X_i) and (w_i) are i.i.d. sequences of random variables, X_i is *uniformly distributed* on $[0,1]^d$ and $Ew_i = 0,\ Ew_i^2 \leq \sigma_w^2$. These assumptions are introduced for the sake of simplicity. They can be weakened, in particular the (unusual) assumption that X is uniformly distributed can easily be relaxed, see [11], this would introduce additional burden to our presentation, however.

For $f \in L_2$, recall the wavelet expansion

$$f(x) \quad = \quad \sum_{k \in Z} \alpha_{0k} \ \Phi_{0k}(x) + \sum_{j=0}^{\infty} \sum_{k \in Z^d} \sum_{l=1}^{2^d-1} \beta_{jk}^{(l)} \ \Psi_{jk}^{(l)}(x) \ , \tag{64}$$

where

$$\alpha_{0k} = \int f(x)\Phi_{0k}(x)dx \quad \text{and} \quad \beta_{jk}^{(l)} = \int f(x)\Psi_{jk}^{(l)}(x)dx \ . \tag{65}$$

To construct an estimate of f a first idea consists in using the law of large numbers and replacing, in expansion (64), the coefficients α_k and $\beta_{jk}^{(l)}$ by their empirical estimates

$$\widehat{\alpha}_{0k}(N) = \frac{1}{N}\sum_{i=1}^{N} Y_i \Phi_{0k}(X_i) \quad \text{and} \quad \widehat{\beta}_{jk}^{(l)}(N) = \frac{1}{N}\sum_{i=1}^{N} Y_i \Psi_{jk}^{(l)}(X_i) \ . \tag{66}$$

Note that the assumption that input X is uniformly distributed has been used at this point.

Density estimation. Assume independent observations $X_1, \ldots, X_N$ of some random variable X with unknown density $f(x)$ are available. Again f can be expanded using (64) (65). But it turns out that

$$\alpha_{0k} = \int f(x)\Phi_{0k}(x)dx \quad = \quad E_f\,\Phi_{0k}(X_i)$$

where E_f denotes expectation with respect to density f, and the same holds for the β's. Thus empirical estimates of the wavelet coefficients α_k and β_{jk} are given by

$$\widehat{\alpha}_{0k} = \frac{1}{N}\sum_{i=1}^{N}\Phi_{0k}(X_i) \quad \text{and} \quad \widehat{\beta}_{jk}^{(l)} = \frac{1}{N}\sum_{i=1}^{N}\Psi_{jk}^{(l)}(X_i)\ . \tag{67}$$

Thus both non-parametric regression and density estimation are faced with the same issue: in formulas (66) and (67), there may not even be X_i's available within the support of many of the Φ's and Ψ's! We shall now discuss this key point for the case of density estimation.

Obviously, in order to compute the empirical coefficient $\widehat{\beta}_{jk}^{(l)}$, we need that at least several observations X_i hit the support of $\Psi_{jk}^{(l)}(x)$. Statistical laws of *loglog* type guarantee that this will generically hold for scales that are not too fine. More specifically, for $j \leq j_{\max}$, where

$$\frac{N}{\ln N} \leq 2^{d\,j_{\max}} \leq \frac{2N}{\ln N}$$

Thus we brute force set $\widehat{\beta}_{jk}^{(l)} = 0$ for $j > j_{\max}$. At this point we have built an estimator of the linear projection type, as in the case of Fourier series in Section 5. Since these estimators are linear, we cannot expect them to be efficient for Besov spaces [37].

A first proposal. Our first attempt to construct an "interesting estimate" is, following the intuition at the end of the previous section, to keep a properly chosen number of coefficients with largest absolute values, and set the others to zero. More precisely, let us consider the set $\widehat{\Lambda}_n$ of pairs $\lambda = (j,k)$ corresponding to the n estimated wavelet coefficients $\widehat{\beta}_{jk}^{(l)}$ with largest absolute values. We construct the estimate $\widehat{f}_N$ as follows:

$$\widehat{f}_N(x) \quad = \quad \underbrace{\sum_k \widehat{\alpha}_{0k}\ \Phi_{0k}(x)}_{\substack{m \text{ coeffs.} \neq 0 \\ (f,\Phi \text{ compact. supp.})}} \quad + \underbrace{\sum_{j=0}^{\infty}\sum_{k\in Z^d}\sum_{l=1}^{2^d-1}\widehat{\beta}_{jk}^{(l)}\ \Psi_{jk}^{(l)}(x)}_{\text{keep the largest } n-m \text{ coeffs.}}\ . \tag{68}$$

The following result can be proved about estimate (68) (see (15) for the definition of the Besov spaces):

Theorem 7 *Let $f \in \mathcal{B}_{p\infty}^s$ with $s \geq d/p$, $\|f\|_\infty < \infty$. If $n = N^{1/(2s+d)}$ is selected in (68), then*

$$E\|\widehat{f}_N - f\|_2^2 = O\left(\frac{\ln N}{N}\right)^{\frac{2s}{2s+d}}. \tag{69}$$

The idea of the proof of Theorem 7 is quite intuitive and typical for wavelet estimators. We follow the argument at the end of the previous section with the only following difference: since no information is available about the distribution of the error $|\widehat{\beta}_\lambda - \beta_\lambda|$ for $\lambda \in \widehat{\Lambda}_n$, we take a cautious upper bound for it:

$$E\;|\widehat{\beta}_\lambda - \beta_\lambda|^2\; 1_{\{\widehat{\beta}\neq 0\}} \quad\leq\quad E\;\sup_{i,j,k} |\widehat{\beta}^{(l)}_{jk} - \beta^{(l)}_{jk}|^2 = O\left(\frac{\ln N}{N}\right),$$

which explains the extra logarithmic factor in (69).

The final solution. Note that n in Theorem 7 depends on s, which is generally unknown. Hence, to complete the estimation algorithm, we need a method to estimate our model order n. Though Generalized Cross-Validation type techniques could be used, we prefer a somewhat different estimation approach developed by D. Donoho, I. Johnstone, G. Kerkyacharian and D. Picard (see the references below). It uses simple thresholding rules [12]:

$$\widetilde{\beta}^{(l)}_{jk} \quad=\quad \widehat{\beta}^{(l)}_{jk}\; 1_{\{|\widehat{\beta}|\geq\lambda_j\}} \tag{70}$$

where λ_j is a threshold parameter, so we set

$$\widehat{f}(x) \quad=\quad \sum_k \widehat{\alpha}_{0k}\;\Phi_{0k}(x) + \sum_{j=0}^{\infty}\sum_{k\in Z^d}\sum_{l=1}^{2^d-1}\widehat{\beta}^{(l)}_{jk}\; 1_{\{|\widehat{\beta}^{(l)}_{jk}|\geq\lambda_j\}}\Psi^{(l)}_{jk}(x)\,. \tag{71}$$

In other words, in expansion (64), we keep those empirical estimates of wavelet coefficients which exceed some properly selected threshold. How this threshold should be selected is provided by the following result:

Theorem 8 ([22] and [23]) *Let $f \in \mathcal{B}^s_{p\infty}$ with $s \geq d/p$, $\|f\|_\infty < \infty$. Select $\lambda_j = \lambda = \sqrt{\frac{C\ln N}{N}}$, with an appropriate $C < \infty$. Then*

$$E\|\widehat{f}_N - f\|_2^2 = O\left(\frac{\ln N}{N}\right)^{\frac{2s}{2s+d}}.$$

The constant C in the expression for the threshold parameter λ is a sort of an "hyperparameter" of the procedure, it can be easily estimated, see [10] and [22] for related discussions. Note that the estimator $\widehat{f}_N$ is adaptive because *it does not require prior knowledge of the regularity parameter.*

Discussion.

- Theorem 8 has the following intuitive explanation. As already mentioned, Besov classes $\mathcal{B}^s_{pq}$ for $p < 2$ have a special structure: a relatively small number of "important" wavelet coefficients are sufficient for obtaining a good function approximation. In the wavelet decomposition $(\widehat{\alpha}_k, \widehat{\beta}^{(l)}_{jk})$ using noisy data, all coefficients are "contaminated" by noise. A Central Limit theorem argument suggests that this noise is approximately Gaussian with zero mean and

[12] we consider here the so called "hard thresholding", meanwhile, other rules can also be studied, for example the "soft thresholding" [19]. See also the discussion in [10].

variance $O(1/N)$. Thus a loglog law implies that the maximal error in the estimates has magnitude given by

$$\max_{j,k}|\widehat{\beta}_{jk}^{(l)} - \beta_{jk}^{(l)}| \approx \sqrt{\frac{2\ln N}{N}}\ .$$

Thus when small (according to threshold λ in Theorem 8) coefficients are shrunk to zero, noise is canceled with very high probability. On the other hand, coefficients exceeding this threshold are likely to be significantly different from zero. This property of thresholding explains another useful feature of the estimator: the estimate $\widehat{f}_N$ has the same regularity as the unknown function f to be estimated (cf. discussion in [23]).

- Let us now consider again our example of estimating the regression function or density $f(x) = 1_{\{0\leq x<a\}}$. Theorem 8 states that the mean square rate of convergence of the wavelet estimator for any bounded function $f \in \mathcal{B}^s_{s^{-1}\infty}$ is very close to $O(N^{-1})$, which is nearly as good as the "parametric" rate of convergence, though the function we estimate is not even continuous. Let us compare the results above with the lower rate of convergence for this problem obtained in [49]. Using the comments 2. of Section 2.2, the following lower bound is a direct corollary of the results of [49] which were originally formulated in terms of Sobolev spaces:

$$\inf_{\widehat{f}_N} \sup_{f\in\mathcal{B}^s_{pq}} E\|\widehat{f}_N - f\|_2 \geq CN^{-2s/(2s+d)} \tag{72}$$

  for any estimator $\widehat{f}_N$. As compared to (72), there is an extra logarithmic factor in the upper bound of Theorem 8. In the more subtle construction presented in [18], this logarithmic factor is eliminated (and even a precise minimax constant is obtained) in the case of Gaussian noises and deterministic design (observations are $x_i = i/N$, $i = 1,\ldots,N$). In [20] a cross-validation procedure is proposed to adapt the optimal algorithm to unknown smoothness. Finally, in [10] the authors of this paper showed that properly selecting the threshold λ for shrinking provides the optimal rate of convergence (without a logarithmic factor). An adaptive version of this algorithm is developed in [36].

- Finally, *a more practical version of this wavelet shrinking algorithm is presented and discussed in [61].* In particular, for the non-parametric regression problem, this version relaxes the irrealistic assumption that input X is uniformly distributed.

7 Application to non-parametric autoregression identification

Let us apply the above developed technique above to the problem of identification of a simple nonlinear dynamic system. Consider the system

$$y_i = f(y_{i-1}) + e_i, \quad y_0 \in R, \quad i = 1,...,N. \tag{73}$$

where (e_i) is a sequence of independent and identically distributed random variables, $Ee_1 = 0$, $Ee_1^2 = \sigma_e^2$. We want to estimate the function $f(x)$. It is clear that the accuracy of this estimate cannot be measured using usual L_p norms on R. Indeed, if the system (73) is stable, then most of the observations are concentrated in a compact set, thus the value $f(x)$ for large x cannot be reconstructed with reasonable precision. However, as we have already seen in section 1, a "common sense" measure of the accuracy (risk) of the estimate $\hat{f}$ of f could be

$$R_N(\widehat{f}, f) \equiv E|\hat{f}_N(y_N) - f(y_N)|^2 \sim \int E \int |\hat{f}_N(x) - f(x)|^2 p_y(x)dx,$$

where p_y is the invariant density of the Markov chain (y_i). We can implement wavelet shrinkage algorithm for this system.

We are confronted here with two problems: the first one is that the observations y_i are dependent; the second one is that the density of observations y_i is not uniform. This first difficulty is handled by fine theoretical considerations without any impact on the estimation algoirithm. To deal with it we require certain stability conditions of the Markov chain (y_i) (see [11] for details). The second problem requires that the optimal threshold in the shrinkage algorithms now depends on the index k of the empiric coefficient $\widehat{\beta}_{jk}$.[13] We can cope with this problem also: it appears that in this case the ideal threshold would be

$$\lambda_{jk} = \sigma_e \sqrt{\frac{2\ln N}{N p_y(2^{-j}k)}}, \tag{74}$$

where p_y is the invariant density of the Markov chain (y_i), i.e., higher threshold is used in low density regions. Though this density is not known *a priori*, it can be easily estimated using, for instance, a simple histogram estimator $\widehat{p}$.

It can be shown that under certain conditions on the underlying Markov chain this procedure supplies us with optimal (in the minimax sense) estimates of f (cf. [11]). To illustrate the efficiency of this method we present here a simulation example: we estimate the function $f(\cdot)$ using the fast wavelet estimator described in [61] with adaptive thresholds $\widehat{\lambda}_{jk}$ estimated according to (74) with $p(\cdot)$ being substituted with $\widehat{p}(\cdot)$. We consider the following three functions f in the model (73):

$$\begin{array}{ll} f(y) = y - 2\text{sign}(y) & \text{"Sigmoidal" signal} \\ f(y) = 0.9y, & \text{"Linear" signal} \\ f(y) = y - 2(\text{sign}(y) - 0.9\sin(y)) & \text{"Sine" signal.} \end{array}$$

(e_i), $i = 1, ..., N$ is supposed to be a sequence of independent Gaussian random variables with $\sigma_e^2 = 1$. The estimate $\widehat{f}$ is computed using $N = 1000$ observations y_i, $i = 1, ..., N$.

On figures 1–3 we present together the signal (y_i), the resulting estimate, and the histogram estimate $\widehat{p}$ of the density $p(y)$.[14] The estimated values of the risk $R_N(\widehat{f}, f)$ are presented on the figures where it is called "error". We consider that the proposed algorithm provides a good quality of visual reconstruction. Amazing quality of fit is obtained for f linear. Indeed, in this case the well known Cramer-Rao lower bound[15] gives

$$R_N(\widehat{f}, f) \geq \frac{\sigma_e^2}{N} = 0.001.$$

One can see that the corresponding value for the wavelet estimator is of the same order of magnitude.

[13] Recall, in the case of uniformly distributed inputs we used the uniform threshold $\lambda_{jk} = \lambda$.

[14] We use the `TeachWave` package by D. Donoho and I. Johnstone [21].

[15] this is a parametric problem!

8 Estimation of high-dimensional systems

As we have seen in Section 3, the curse of dimensionality is encountered when estimating functions with large dimensional inputs, and one has to resort to methods that reduce the effective dimension of the input. All the methods reported in Section 3 can be, and have been (sometimes successfully) used in estimation. However, there are only a few theoretical results available today. In this section we will show how the approximation technique given in Section 3.4 can be adapted to the estimation problem.

8.1 Wavelets

We will construct now an estimation algorithm based on the approximation result of Theorem 4. It is essentially based on the idea of truncation proposed in [19].

Estimation algorithm :

1. **sample length N, set $\beta = d - 1/2$ and**

$$n = \sqrt{N}, \quad a_{\min} = N^{-\frac{1}{2d}}, \quad a_{\max} = (\frac{N}{\ln N})^{1/d}$$

2. **compute**

$$\begin{aligned} \widehat{u}_N(a,t) &= \frac{b^d}{N}\sum_{i=1}^{N} y_i \varphi(a(x_i - t))1_{\{a_{\min}\le a \le a_{\max}\}} \\ \widehat{v}_N(a,t) &= = a^{\beta}\widehat{u}_N(a,t) \end{aligned}$$

3. **and truncate**

$$\tilde{v}_N(a,t) = \begin{cases} \widehat{v}_N(a,t) & \text{if } |\widehat{v}_N(a,t)| \ge K\sqrt{\frac{\ln N}{N}}; \\ 0 & \text{if } |\widehat{v}_N(a,t)| < K\sqrt{\frac{\ln N}{N}} \end{cases} \tag{75}$$

4. **compute $C_N = \|a^{(d-1)/2}\tilde{v}_N(a,t)\|_1$ and draw n independent points $(a_1,t_1),\ldots,(a_n,t_n)$ according to the density $\tilde{w}_N(a,t) = |a^{(d-1)/2}\tilde{v}_N(a,t)|/C_N$;**

5. **compute**

$$\widehat{f}_N(x) = C_N n^{-1}\sum_{k=1}^{n}(a_k)^{d/2}\ \mathrm{sign}(\widehat{u}(a_k,t_k))\ \psi(a_k(x-t_k)) \tag{76}$$

The key point of the algorithm is the truncation procedure of point **3** in which we shrink to 0 the "density" $\widehat{v}_N$ if its amplitude is less than $K\sqrt{\ln N/N}$. This step has a nice heuristic explanation (cf. Section 6.2). Note that $\widehat{v}_N(a,t)$ can be decomposed as $\widehat{v}_N(a,t) = v(a,t) + \varepsilon_N(a,t)$, where $v_N(\cdot)$ is the corresponding true density, and $\varepsilon(\cdot)$ is a centered random noise with variance of order N^{-1}. On the other hand, the condition $\|w\|_1 = \|a^{(d-1)/2}v\|_1 < \infty$ implies that $v(a,t) = o(1/\sqrt{N})$ on a subset of $\{a \le (\frac{N}{\ln N})^{1/d}\}$ of the underlying measure. So, putting all these terms to the sum (76) would mean adding noise to our estimate and negligible useful information. Therefore, we can improve significantly our estimate when shrinking these terms to the zero.

The only parameter of the algorithm to be chosen is the truncation parameter K in (75). It depends on $\|f\|_\infty$ and on Ee_1^2, and can be estimated "on line". Furthermore, the algorithm is not too sensitive to this parameter, and it can be often chosen on the basis of the available a priori information.

The theoretic analysis reveals that if the function $f(\cdot)$ is bounded and $f \in \mathcal{W}_1^{d/2+\varepsilon} \cap \mathcal{W}_2^{d/4+\varepsilon}$ for some $\varepsilon > 0$, then the following bound holds for the estimate $\widehat{f}_N$:

$$E\|\widehat{f}_N - f\|_2^2 \le C \left(\frac{\ln N}{N}\right)^{1/2}$$

for some $C < \infty$. This means that the proposed algorithm achieves "up to ε" the rate of convergence of the estimators discussed in Section 6.2.

In [12] another truncation algorithm is proposed which allows us to simplify significantly the very unpleasant part of the algorithm – the drawing independent samples from the density $\tilde{w}_N$ (line 4 of the algorithm).

9 Conclusion : the gap between theory and everyday practice

In this paper, we have surveyed and discussed part of the mathematical foundations of nonlinear black-box modeling. What mathematics tells us can be summarized as follows :

- The bias/variance trade-off is a key factor, as it is always in system identification when model order is not known a priori. Error variance depends on how many parameters are used for fitting. Thus efforts concentrate on reducing the bias without increasing the number of parameters in the model. Since we are dealing with nonlinear systems, there is much more flexibility in handling this problem than in linear system identification. Thus we have paid a lot of attention to *function approximation* issues as a prior topic before considering estimation from noisy data.

- Function approximation very much depends on the function space in consideration. Thus we are faced with the new problem of having to specify the function space which our unknown system is supposed to belong to, this is prior information referred to as the *smoothness class.* Such kind of prior information is not easily at hand, however, thus it is of primary importance to design procedures which are equally efficient for various classes of functions.

- Approximation (and estimation) methods have been known for a long time, which perform equally well for various smoothness classes of systems, provided that they consist of *uniformly smooth* systems. Among such methods one finds classics of non-parametric statistics such as kernel or linear projection estimates.

- These methods perform poorly on systems that are smooth, but with some spikes and jumps. Unfortunately, such nonlinear systems are frequently encountered in practice. Approximation and estimation methods have been proposed, which are *spatially adaptive,* i.e., which are able to locally adapt the smoothness of the approximants or estimates to the function to be approximated. They proved efficient and successful in practice. They were more or less attractive, depending on their computational and memory cost (as well as psychological appeal). Within this large army of methods, neural networks reached the top in celebrity. Results are available which mathematically support the success of these methods for *approximation,* but few results are available to support their use for *estimation.*

- *Besov spaces* revealed to be an adequate parameterized family of function spaces to model spikyness and jumpiness in a tunable way. They support most of the mathematical results about approximation and convergence rates.

- *Wavelets* revealed to be nicely associated with Besov spaces since Besov norms are easily evaluated using wavelet decompositions. Thus wavelet based estimation algorithms are the only class of algorithms for which complete analysis is available today, both for approximation and estimation. These theoretical results show optimality of these algorithms.

- Getting *lower bounds* for convergence rates of estimation procedures is much more involved than for parametric estimation, in which key tools are Cramer-Rao bound and Fisher information. Lower rates and minimax optimality were introduced to this end. These tools are much more technical and difficult to use than Cramer-Rao bound and Fisher information.

- The *curse of dimensionality* refers to the fact that one barely has enough data for fitting when the input dimension is large. The notion of "effective dimension" which we discussed somewhat informally plays an important role. Function classes of low "effective dimension" can be found and used for analysis in the case of large dimensional inputs.

- Altogether, non-parametric estimation can also be considered as maximum likelihood estimation for unknown systems with unknown and unbounded model order. Consistency, convergence rates, and AIC/BIC/.../XIC criteria are the classics. Two difficulties occur in applying this point of view to black-box nonlinear system identification. First, the "good" parameter dimension is typically very large so that AIC/BIC/.../XIC are not very practical. Second, and most important from the theoretical point of view, asymptotic results involve the argument of the minimum of the likelihood. But the likelihood is very non-convex and no procedure is provably known to find the minimum. This is why we did not pay a large attention to this point of view in this paper, where emphasis was on mathematical results only.

Obviously, the everyday practice in nonlinear black-box modeling is quite different from the direct implementation of mathematical advises. Now, there are several factors which may explain the gap between the mathematical foundations and the practical confidence in this or that method, in particular :

- Computational cost and memory requirements could be formally considered, but have not been discussed in this paper.

- How bad is a non convex functional for optimization, especially in large dimension, is hard to assess.

Since authors of this article and the companion one [61] are the same, they would be quite schizophrenic if they would blindly and only trust the mathematics as a guideline for implementation. Nonetheless, mathematical results have the advantage of providing theoretically sound intuition and guidelines about how and why methods perform good or bad in various situations.

References

[1] H. Akaike, *Statistical predictor identification*, Ann. Inst. Math. Statist., 22 (1970), pp. 203–217.

[2] A. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. on Information Theory, 39 (1993).

[3] R. Bellman, *Adaptive Control Processes*, Princetion University Press, Princeton, 1966.

[4] O. Besov, *On a family of functional spaces : embedding theorems and applications*, Doklady Acad. Nauk SSSR, 126, 1163-1165 (1959).

[5] L. Breiman, *Hinging hyperplanes for regression, classification and function approximation*, IEEE Trans. on Information Theory, 39 (1993), pp. 999–1013.

[6] L. Breiman, J. Friedman, J. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, Belmont, California, 1984.

[7] N. Cencov, *Statistical decision rules and optimal inference*, Amer. Math. Soc. Transl., 53 (1982). Providence, R.I.

[8] P. Craven and G. Wahba, *Smoothing noisy data with spline functions*, Numer. Math., 31 (1979), pp. 337–403.

[9] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF regional series in applied mathematics, CBMS-NSF regional conference, 1992.

[10] B. Delyon and A. Juditsky, *Wavelet Estimators, Global Error Measures Revisited*, Technical Report 782, IRISA, 1993.

[11] ——, *Optimal Estimators for Functional Autoregression*, Tech. Rep. IRISA, in preparation, 1995.

[12] B. Delyon, A. Juditsky, and A. Benveniste, *Accuracy analysis for wavelet networks*, IEEE Transactions on neural networks, (1994). to appear.

[13] R. DeVore, B. Jawerth, and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math., To appear (1994).

[14] L. Devroye, *Any discrimination rule can have an arbitrary bad probability of error for final sample size*, IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-4 (1982), pp. 154–157.

[15] L. Devroye and L. Györfi, *Nonparametric Density Estimation L_1 View*, J. Wiley, New-York, 1985.

[16] L. Devroye and T. Wagner, *Distribution free consistency result in nonparametric discrimination and regression function estimation*, The Annals of Statistics, 8 (1980), pp. 231–239.

[17] D. Donoho and I. Johnstone, *Projection-based approximation and a duality with kernel methods*, Ann. Statist., 17:58–106 (1989).

[18] ——, *Minimax Estimation via Wavelet Shrinkage*, Technical Report, Department of Statistics, Stanford University `ftp playfair.stanford.edu`, 1992.

[19] ——, *Minimax Risk over l_p-balls*, Technical Report, Department of Statistics, Stanford University `ftp playfair.stanford.edu`, 1992.

[20] ——, *Adapting to Unknown Smoothness via Wavelet Shrinkage*, Technical Report, Department of Statistics, Stanford University, `ftp playfair.stanford.edu`, 1993.

[21] ——, *TeachWave V 0.550*, Department of Statistics, Stanford University `wavelab@playfair.stanford.edu`, 1993.

[22] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, *Density estimation by wavelet thresholding*, Technical Report, Department of Statistics, Stanford University `ftp playfair.stanford.edu`, 1993.

[23] ——, *Wavelet Shrinkage: Asymptopia*, Manuscript on `ftp playfair.stanford.edu`, 1993.

[24] M. Duflo, *Recursive Stochastic Methods*, Springer-Verlag, Berlin, 1993.

[25] S. Efroimovich and M. Pinsker, *Estimation of square-integrable spectral density based on a sequence of observations*, Problems of Information Transmission (in Russian), pp. 182-196 (1982).

[26] ——, *Estimation of square-integrable probability density of a random variable*, Problems of Information Transmission (in Russian), pp. 175-189 (1983).

[27] ——, *A learning algorithm for nonparametric filtering*, Automatika i Telemehanika (in Russian), 11, 58-65 (1984).

[28] J. Fan and I. Gijbels, *Variable bandwidth adn local linear regression smoothers*, Annals of Statistics, 20 (1992), pp. 2008–2036.

[29] J. Friedman, *Multivariate adaptive regression splines (with discussion)*, The Annals of Statistics, 19 (1991), pp. 1–141.

[30] J. Friedman and W. Stuetzle, *Projection pursuit regression*, J. Amer. Stat. Assoc., 76 (1981), pp. 817–823.

[31] W. Härdle and J. Marron, *Optimal bandwidth selection in nonparametric regression function estimation*, The Annals of Statistics, 13 (1985), pp. 1465–1481.

[32] P. Huber, *Projection pursuit (with discussion)*, The Annals of Statistics, 13 (1985), pp. 435–475.

[33] K. Hunt, D. Sbarbaro, R. Zbikowski, and P. Gawthrop, *Neural networks for control systems — a survey*, Automatica, 28 n°6 (1992), pp. 1083–1112.

[34] I. Ibragimov and R. Khasminskij, *Statistical Estimation Asymptotic Theory*, Springer-Verlag, Berlin, 1981.

[35] S. Jaffard and P. Laurentçot, *Wavelets : A Tutorial*, Academic Press, 1989, ch. Wavelets and P.D.E.'s.

[36] A. Juditsky, *Adaptive Wavelet Estimators*, Technical Report 815, IRISA, 1994.

[37] G. Kerkyacharian and D. Picard, *Density estimation in besov spaces*, Stat. and Prob. Letters, 13, 15-24 (1992).

[38] A. Kolmogorov, *On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition*, Dokl, 114 (1957), pp. 679–681.

[39] A. Korostelev and A. Tsybakov, *Minimax Theory of Image Reconstruction*, Springer-Verlag, Berlin, 1981.

[40] K. Li, *Asymptotic optimality of c_L and generalized cross-validation in ridge regression and application to the spline smoothing*, The Annals of Statistics, 14 (1986), pp. 1101–1112.

[41] ——, *Asymptotic optimality of c_L and generalized cross-validation : discrete index set*, The Annals of Statistics, 15 (1987), pp. 958–975.

[42] L. Ljung, *System Identification : Theory for the User*, Prentice Hall Information and System Sciences Series, Prentice Hall, 1987.

[43] G. Lorentz, *The 13th problem of hilbert*, in Mathematical Developments Arising from Hilbert Problems, F. Browder, ed., Am. MAth. Soc., 1976, p. .

[44] C. Mallows, *Statistical predictor identification*, Technometrics, 15 (1973), pp. 661–675.

[45] Y. Meyer, *Ondelettes et Opérateurs*, Hermann, 1990.

[46] J. Morgan and J. Sonquist, *Problems in the analysis of survey data, and a proposal*, J. Amer. Stat. Assoc., 58 (1963), pp. 415–434.

[47] H. Müller and U. Stadtmüller, *Variable bandwidth kernel estimators of regression curves*, The Annals of Statistics, 15(1), 182–201 (1987).

[48] E. Nadaraya, *On estimating regression*, Theory of Prob. and Appl., 9 (1964), pp. 141–142.

[49] A. Nemirovskij, *Nonparametric estimation of smooth regression functions*, Izv. Acad. Nauk SSSR, Techn. Kibern. (in Russian), 3, 50-60 (1985).

[50] G. Oppenheim and B. Portier, *Commande adaptative du processus de Markov $x_{t+1} = f_t + u_t + x_t,\ \ t \in N$*, Technical Report 90–18, Université d'Orsay, 1990.

[51] E. Parzen, *On estimation of probability density function and the mode*, Ann. of Math. Stat., 33 (1962), pp. 1065–1076.

[52] P. Petrushev and V. Popov, *Rational Approximation Of Real Functions*, Cambridge University Press, Cambridge, 1987.

[53] A. Pinkus, *n-Widths in Approximation Theory*, Springer-Verlag, Berlin, 1985.

[54] T. Poggio and F. Girosi, *Networks for approximation and learning*, Proceedings of the IEEE, 78 (1990), pp. 1481–1497.

[55] B. Polyak and A. Tsybakov, *Asymptotical optimality of c_p criterion for projection regression estimates*, Theory of Prob. and Appl., 35 (1990), pp. 305–317.

[56] B. Portier, *Estimation non paramétrique et commande adaptative de processus Markoviens non linéaires*, PhD thesis, Université Paris Sud, Orsay, 1992.

[57] J. Rice, *Bandwidth choice for nonparametric regression*, The Annals of Statistics, 12 (1984), pp. 1215–1230.

[58] M. Rosenblatt, *Remarks on some nonparametric estimates of density functions*, Ann. of Math. Stat., 27 (1956), pp. 832–835.

[59] ——, *Curve estimation*, Ann. of Math. Stat., 42 (1971), pp. 1815–1842.

[60] W. Sickel, *Spline representations of functions in besov-triebel-lizorkin spaces on R^n*, Forum Math., 2, 451-476 (1990).

[61] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. G. H. Hjalmarsson, and A. Juditsky, *Nonlinear black-box modeling in system identification: a unified overview*, Automatica, (1995), p. to appear.

[62] T. Söderström and P. Stoica, *System Identification*, Prentice-Hall International, Hemel Hempstead, Hertfordshire, 1989.

[63] E. Sontag, *Nonlinear regulation : the piecewise linear approach*, IEEE Trans. on Automatic Control, 26 (1981), pp. 346–358.

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

[65] H. Triebel, *Theory of Function Spaces*, Birkhäuser Verlag, Berlin, 1983.

[66] ——, *Theory of Function Spaces II*, Birkhäuser Verlag, Berlin, 1993.

[67] P. Vieu, *Nonparametric regression: optimal local bandwidth choice*, J.R. Statist. Soc. Ser. B, 53 (1991), pp. 453–464.

[68] G. Wahba, *Spline functions for observational data*, SIAM, Philadelphia, PA, 1991.

[69] G. Watson, *Smooth regression analysis*, Sankhya, Series, A (1969), pp. 359–372.

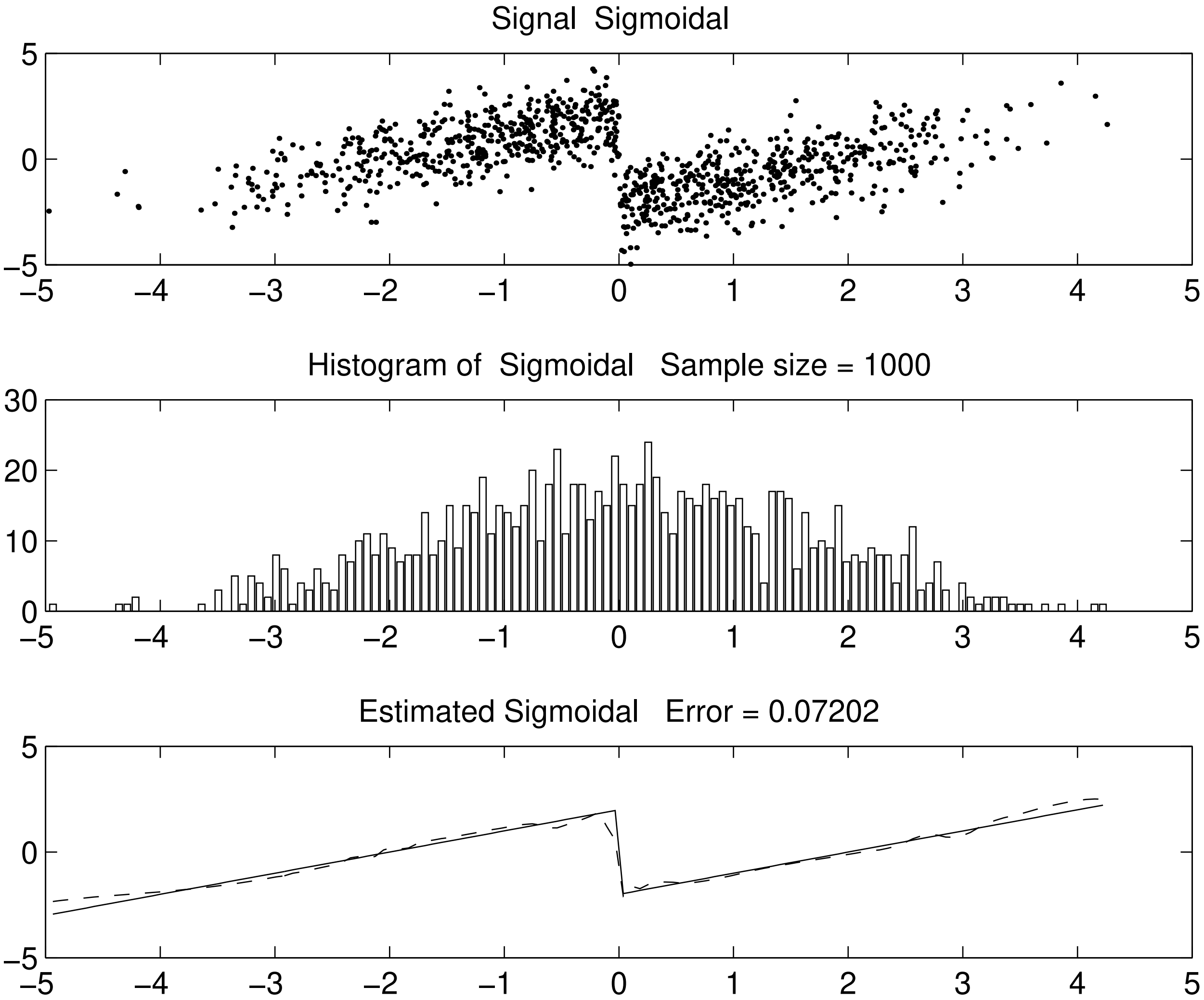
Signal Sigmoidal
Histogram of Sigmoidal Sample size = 1000
Estimated Sigmoidal Error = 0.07202

Figure 1:

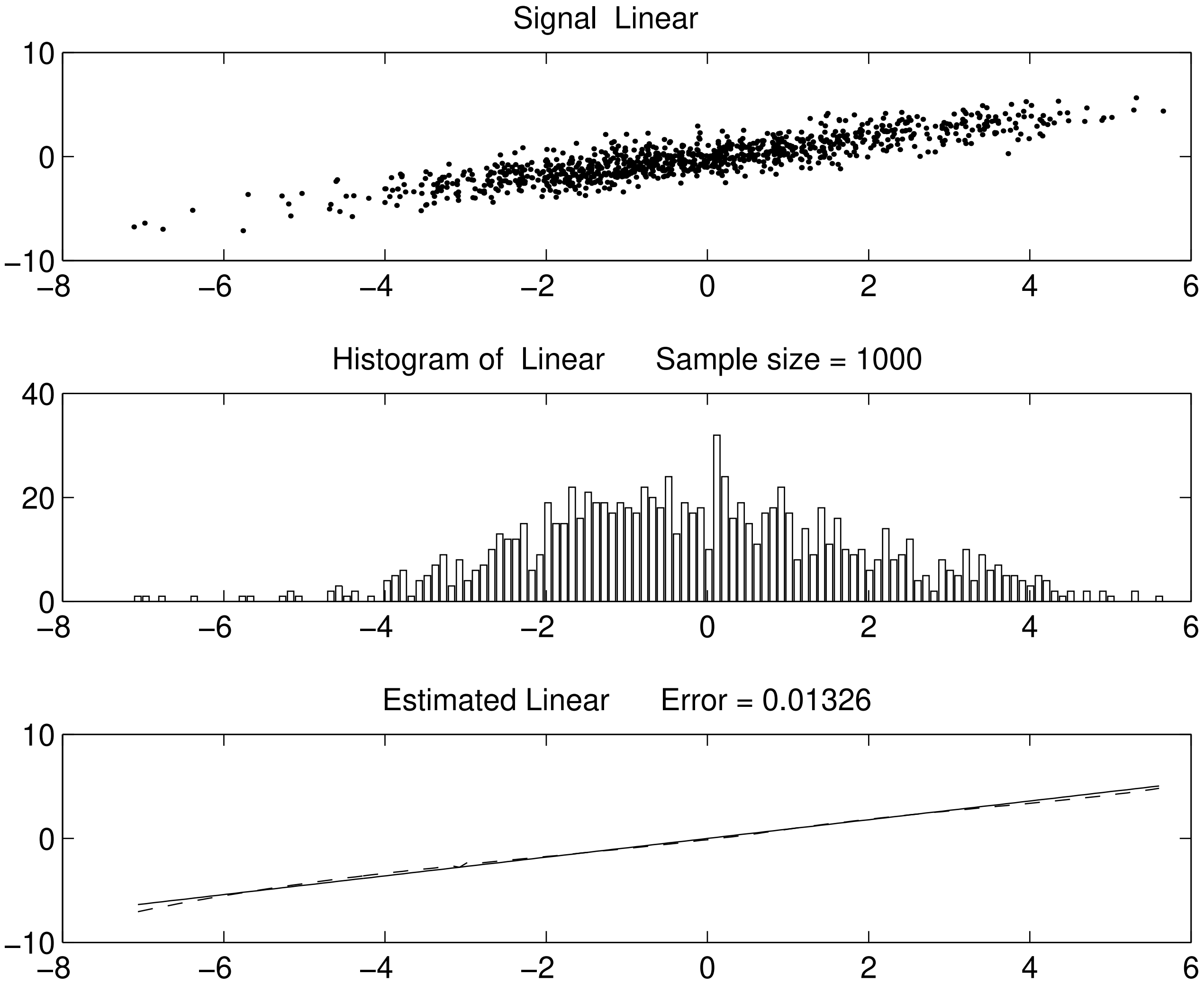
Signal Linear
10
0
−10
−8
−6
−4
−2
0
2
4
6
Histogram of Linear   Sample size = 1000
40
20
0
Estimated Linear   Error = 0.01326

Figure 2:

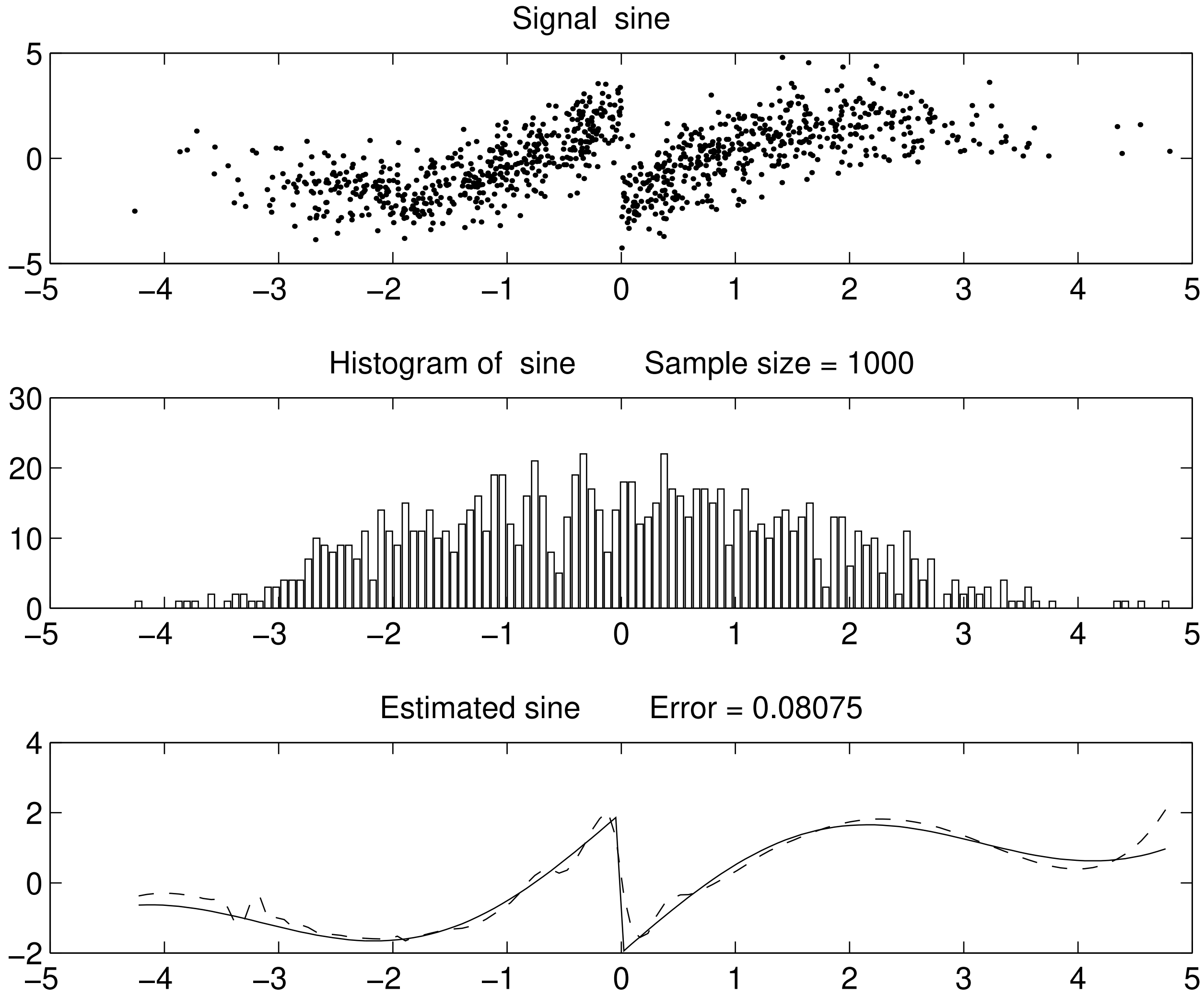
Signal sine
Histogram of sine Sample size = 1000
Estimated sine Error = 0.08075

Figure 3: