Acta Numerica (1998), pp. 51–150

Nonlinear approximation

Ronald A. DeVore*
Department of Mathematics,
University of South Carolina,
Columbia, SC 29208, USA
E-mail: devore@math.sc.edu

This is a survey of nonlinear approximation, especially that part of the subject which is important in numerical computation. Nonlinear approximation means that the approximants do not come from linear spaces but rather from nonlinear manifolds. The central question to be studied is what, if any, are the advantages of nonlinear approximation over the simpler, more established, linear methods. This question is answered by studying the rate of approximation which is the decrease in error versus the number of parameters in the approximant. The number of parameters usually correlates well with computational effort. It is shown that in many settings the rate of nonlinear approximation can be characterized by certain smoothness conditions which are significantly weaker than required in the linear theory. Emphasis in the survey will be placed on approximation by piecewise polynomials and wavelets as well as their numerical implementation. Results on highly nonlinear methods such as optimal basis selection and greedy algorithms (adaptive pursuit) are also given. Applications to image processing, statistical estimation, regularity for PDEs, and adaptive algorithms are discussed.

* This research was supported by Office of Naval Research Contract N0014-91-J1343 and Army Research Office Contract N00014-97-1-0806.

CONTENTS

1. Nonlinear approximation: an overview

The fundamental problem of approximation theory is to resolve a possibly complicated function, called the *target function*, by simpler, easier to compute functions called the *approximants*. Increasing the resolution of the target function can generally only be achieved by increasing the complexity of the approximants. The understanding of this trade-off between resolution and complexity is the main goal of constructive approximation. Thus the goals of approximation theory and numerical computation are similar, even though approximation theory is less concerned with computational issues. The differing point in the two subjects lies in the information assumed to be known about the target function. In approximation theory, one usually assumes that the values of certain simple linear functionals applied to the target function are known. This information is then used to construct an approximant. In numerical computation, information usually comes in a different, less explicit form. For example, the target function may be the solution to an integral equation or boundary value problem and the numerical analyst needs to translate this into more direct information about the target function. Nevertheless, the two subjects of approximation and computation are inexorably intertwined and it is impossible to understand fully the possibilities in numerical computation without a good understanding of the elements of constructive approximation.

It is noteworthy that the developments of approximation theory and numerical computation followed roughly the same line. The early methods utilized approximation from finite-dimensional linear spaces. In the beginning, these were typically spaces of polynomials, both algebraic and trigonometric. The fundamental problems concerning order of approximation were solved in this setting (primarily by the Russian school of Bernstein, Chebyshev, and their mathematical descendants). Then, starting in the late 1950s came the development of piecewise polynomials and splines and their incorporation into numerical computation. We have in mind the finite element methods (FEM) and their counterparts in other areas such as numerical quadrature, and statistical estimation.

It was noted shortly thereafter that there was some advantage to be gained by not limiting the approximations to come from linear spaces, and therein emerged the beginnings of nonlinear approximation. Most notable in this regard was the pioneering work of Birman and Solomyak (1967) on adaptive approximation. In this theory, the approximants are not restricted to come from spaces of piecewise polynomials with a fixed partition; rather, the partition was allowed to depend on the target function. However, the number of pieces in the approximant is controlled. This provides a good match with numerical computation since it often represents closely the cost of computation (number of operations). In principle, the idea was simple: we should use a finer mesh where the target function is not very smooth (singular) and a coarser mesh where it is smooth. The paramount question remained, however, as to just how we should measure this smoothness in order to obtain definitive results.

As is often the case, there came a scramble to understand the advantages of this new form of computation (approximation) and, indeed, rather exotic spaces of functions were created (Brudnyi 1974, Bergh and Peetre 1974), to define these advantages. But to most, the theory that emerged seemed too much a tautology and the spaces were not easily understood in terms of classical smoothness (derivatives and differences). But then came the remarkable discovery of Petrushev (1988) (preceded by results of Brudnyi (1974) and Oswald (1980)) that the efficiency of nonlinear spline approximation could be characterized (at least in one variable) by classical smoothness (Besov spaces). Thus the advantage of nonlinear approximation became crystal clear (as we shall explain later in this article).

Another remarkable development came in the 1980s with the development of multilevel techniques. Thus, there were the roughly parallel developments of multigrid theory for integral and differential equations, wavelet analysis in the vein of harmonic analysis and approximation theory, and multiscale filterbanks in the context of image processing. From the viewpoint of approximation theory and harmonic analysis, the wavelet theory was important on several counts. It gave simple and elegant unconditional bases (wavelet bases) for function spaces (Lebesgue, Hardy, Sobolev, Besov, Triebel–Lizorkin) that simplified some aspects of Littlewood–Paley theory (see Meyer (1990)). It provided a very suitable vehicle for the analysis of the core linear operators of harmonic analysis and partial differential equations (Calderón–Zygmund theory). Moreover, it allowed the solution of various functional analytic and statistical extremal problems to be made directly from wavelet coefficients.

Wavelet theory provides simple and powerful decompositions of the target function into a series of building blocks. It is natural, then, to approximate the target function by selecting terms of this series. If we take partial sums of this series we are approximating again from linear spaces. It was easy to establish that this form of linear approximation offered little, if any, advantage over the already well established spline methods. However, it is also possible to let the selection of terms to be chosen from the wavelet series depend on the target function f and keep control only over the number of terms to be used. This is a form of nonlinear approximation which is called *n-term approximation*. This type of approximation was introduced by Schmidt (1907). The idea of n-term approximation was first utilized for multivariate splines by Oskolkov (1979).

Most function norms can be described in terms of wavelet coefficients. Using these descriptions not only simplifies the characterization of functions with a specified approximation order but also makes transparent strategies for achieving good or best n-term approximations. Indeed, it is enough to retain the n terms in the wavelet expansion of the target function that are largest relative to the norm measuring the error of approximation. Viewed in another way, it is enough to threshold the properly normalized wavelet coefficients. This leads to approximation strategies based on what is called *wavelet shrinkage* by Donoho and Johnstone (1994). Wavelet shrinkage is used by these two authors and others to solve several extremal problems in statistical estimation, such as the recovery of the target function in the presence of noise.

Because of the simplicity in describing n-term wavelet approximation, it is natural to try to incorporate a good choice of basis into the approximation problem. This leads to a double stage nonlinear approximation problem where the target function is used both to choose a good (or best) basis from a given library of bases and then to choose the best n-term approximation relative to the good basis. This is a form of *highly nonlinear approximation*. Other examples are greedy algorithms and adaptive pursuit for finding an n-term approximation from a redundant set of functions. Our understanding of these highly nonlinear methods is quite fragmentary. Describing the functions that have a specified rate of approximation with respect to highly nonlinear methods remains a challenging problem.

Our goal in this paper is to be tutorial rather than complete in our description of nonlinear approximation. We spare the reader some of the finer aspects of the subject in search of clarity. In this vein, we begin in Section 2 by considering approximation in a Hilbert space. In this simple setting the problems of linear and nonlinear approximation are easily settled and the distinction between the two subjects is readily seen.

In Section 3, we consider approximation of univariate functions by piecewise constants. This form of approximation is the prototype of both spline approximation and wavelets. Understanding linear and nonlinear approximation by piecewise constants will make the transition to the fuller aspects of splines (Section 6) and wavelets (Section 7) more digestible.

In Section 8, we treat highly nonlinear methods. Results in this subject are in their infancy. Nevertheless, the methods are already in serious numerical use, especially in image processing.

As noted earlier, the thread that runs through this paper is the following question: what properties of a function determine its rate of approximation by a given nonlinear method? The final solution of this problem, when it is known for a specific method of approximation, is most often in terms of Besov spaces. However, we try to postpone the full impact of Besov spaces until the reader has, we hope, developed significant feeling for smoothness conditions and their role in approximation. Nevertheless, it is impossible to understand this subject fully without finally coming to grips with Besov spaces. Fortunately, they are not too difficult when viewed via moduli of smoothness (Section 4) or wavelet coefficients (Section 7).

Nonlinear approximation is used significantly in many applications. Perhaps the greatest success for this subject has been in image processing. Nonlinear approximation explains the thresholding and quantization strategies used in compression and noise removal. It also explains how quantization and thresholding may be altered to accommodate other measures of error. It is also noteworthy that it explains precisely which images can be compressed well by certain thresholding and quantization strategies. We discuss some applications of nonlinear methods to image processing in Section 10.

Another important application of nonlinear approximation lies in the solution of operator equations. Most notable, of course, are the adaptive finite element methods for elliptic equations (see Babuška and Suri (1994)) as well as the emerging nonlinear wavelet methods in the same subject (see Dahmen (1997)). For hyperbolic problems, we have the analogous developments of moving grid methods. Applications of nonlinear approximation in PDEs are touched upon in Section 10.

In approximation theory, one measures the complexity of the approximation process by the number of parameters needed to specify the approximant. This agrees in principle with the concepts of complexity in information theory. However, it does not necessarily agree with computational complexity, which measures the number of computations necessary to render the approximant. This is particularly the case when the target function is not explicitly available and must be computed through a numerical process such as in the numerical solution of integral or differential equations. We shall not touch on this finer notion of computational complexity in this survey. Good references for computational complexity in the framework of linear and nonlinear approximation is given in the book of Traub, Wasilkowski and Woźniakowski (1988), the paper of E. Novak (1996), and the references therein.

Finally, we close this introduction with a couple of helpful remarks about notation. Constants appearing in inequalities will be denoted by C and may vary at each occurrence, even in the same formula. Sometimes we will indicate the parameters on which the constant depends. For example, $C(p)$ (respectively, $C(p, \alpha)$) means the constant depends only on p (respectively, p and α). However, usually the reader will have to consult the text to understand the parameters on which C depends. More ubiquitous is the notation

$$A \asymp B, \tag{1.1}$$

which means there are constants $C_1, C_2 > 0$ such that $C_1 A \leq B \leq C_2 A$. Here A and B are two expressions depending on other variables (parameters). When there is any chance of confusion, we will indicate in the text the parameters on which C_1 and C_2 depend.

2. Approximation in a Hilbert space

The problems of approximation theory are simplest when they take place in a Hilbert space $\mathcal{H}$. Yet the results in this case are not only illuminating but very useful in applications. It is worthwhile, therefore, to begin with a brief discussion of linear and nonlinear approximation in this setting.

Let $\mathcal{H}$ be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|_H$ and let η_k, $k = 1, 2, \ldots$, be an orthonormal basis for $\mathcal{H}$. We shall consider two types of approximation corresponding to the linear and nonlinear settings.

For linear approximation, we use the linear space $\mathcal{H}_n := \text{span}\{\eta_k : 1 \leq k \leq n\}$ to approximate an element $f \in \mathcal{H}$. We measure the approximation error by

$$E_n(f)_{\mathcal{H}} := \inf_{g \in \mathcal{H}_n} \|f - g\|_{\mathcal{H}}. \tag{2.1}$$

As a counterpart in nonlinear approximation, we have *n-term approximation*, which replaces $\mathcal{H}_n$ by the space Σ_n consisting of all elements $g \in \mathcal{H}$ that can be expressed as

$$g = \sum_{k \in \Lambda} c_k \eta_k, \tag{2.2}$$

where $\Lambda \subset \mathbb{N}$ is a set of indices with $\#\Lambda \leq n$.[1] Notice that, in contrast to $\mathcal{H}_n$, the space Σ_n is not linear. A sum of two elements in Σ_n will in general

[1] We use $\mathbb{N}$ to denote the set of natural numbers and $\#S$ to denote the cardinality of a finite set S.

need $2n$ terms in its representation by the η_k. Analogous to E_n, we have the *error of n-term approximation*

$$\sigma_n(f)_{\mathcal{H}} := \inf_{g\in\Sigma_n} \|f-g\|_{\mathcal{H}}. \tag{2.3}$$

We pose the following question. Given a real number $\alpha > 0$, for which elements $f \in \mathcal{H}$ do we have

$$E_n(f)_{\mathcal{H}} \le M n^{-\alpha}, \quad n = 1, 2, \ldots, \tag{2.4}$$

for some constant $M > 0$? Let us denote this class of f by $\mathcal{A}^\alpha((\mathcal{H}_n))$, where our notation reflects the dependence on the sequence $(\mathcal{H}_n)$, and define $|f|_{\mathcal{A}^\alpha((\mathcal{H}_n))}$ as the infimum of all M for which (2.4) holds. $\mathcal{A}^\alpha$ is called an *approximation space*: it gathers under one roof all $f \in \mathcal{H}$ which have a common approximation order. We denote the corresponding class for (Σ_n) by $\mathcal{A}^\alpha((\Sigma_n))$.

We shall see that it is easy to describe the above approximation classes in terms of the coefficients in the orthogonal expansion

$$f = \sum_{k=1}^{\infty} \langle f, \eta_k\rangle \eta_k. \tag{2.5}$$

Let us use in this section the abbreviated notation

$$f_k := \langle f, \eta_k\rangle, \quad k = 1, 2, \ldots. \tag{2.6}$$

Consider first the case of linear approximation. The best approximation to f from $\mathcal{H}_n$ is given by the projection

$$P_n f := \sum_{k=1}^{n} f_k \eta_k \tag{2.7}$$

onto $\mathcal{H}_n$ and the approximation error satisfies

$$E_n(f)^2_{\mathcal{H}} = \sum_{k=n+1}^{\infty} |f_k|^2. \tag{2.8}$$

We can characterize $\mathcal{A}^\alpha$ in terms of the dyadic sums

$$F_m := \left(\sum_{k=2^{m-1}+1}^{2^m} |f_k|^2 \right)^{1/2}, \quad m = 1, 2, \ldots. \tag{2.9}$$

Indeed, it is almost a triviality to see that $f \in \mathcal{A}^\alpha((\mathcal{H}_n))$ if and only if

$$F_m \le M 2^{-m\alpha}, \quad m = 1, 2, \ldots, \tag{2.10}$$

and the smallest M for (2.10) is equivalent to $|f|_{\mathcal{A}^\alpha((\mathcal{H}_n))}$. To some, (2.10) may not seem so pleasing since it is so close to a tautology. However, it usually serves to characterize the approximation spaces $\mathcal{A}_\alpha((\mathcal{H}_n))$ in concrete settings.

It is more enlightening to consider a variant of $\mathcal{A}^\alpha$. Let $\mathcal{A}_2^\alpha((\mathcal{H}_n))$ denote the set of all f such that

$$|f|_{\mathcal{A}_2^\alpha((\mathcal{H}_n))} := \left(\sum_{n=1}^{\infty}[n^\alpha E_n(f)_{\mathcal{H}}]^2 \frac{1}{n}\right)^{1/2} \tag{2.11}$$

is finite. From the monotonicity of $E_k(f)_{\mathcal{H}}$, it follows that

$$|f|_{\mathcal{A}_2^\alpha((\mathcal{H}_n))} \asymp \left(\sum_{k=0}^{\infty} 2^{2k\alpha} E_{2^k}(f)_{\mathcal{H}}^2\right)^{1/2}. \tag{2.12}$$

The condition for membership in $\mathcal{A}_2^\alpha$ is slightly stronger than membership in $\mathcal{A}^\alpha$. The latter requires that the sequence $(n^\alpha E_n)$ is bounded while the former requires that it is square summable with weight $1/n$.

The space $\mathcal{A}_2^\alpha((\mathcal{H}_n))$ is characterized by

$$\sum_{k=1}^{\infty} k^{2\alpha}|f_k|^2 \le M^2 \tag{2.13}$$

and the smallest M satisfying (2.13) is equivalent to $|f|_{\mathcal{A}_2^\alpha((\mathcal{H}_n))}$. We shall give the simple proof of this fact since the ideas in the proof are used often. First of all, note that (2.13) is equivalent to

$$\sum_{m=1}^{\infty} 2^{2m\alpha} F_m^2 \le (M')^2 \tag{2.14}$$

with M of (2.13) and M' of (2.14) comparable. Now, we have

$$2^{2m\alpha} F_m^2 \le 2^{2m\alpha} E_{2^{m-1}}(f)_{\mathcal{H}}^2,$$

which, when using (2.12), gives one of the implications of the asserted equivalence. On the other hand,

$$2^{2m\alpha} E_{2^m}(f)_{\mathcal{H}}^2 = 2^{2m\alpha} \sum_{k=m+1}^{\infty} F_k^2$$

and therefore

$$\sum_{m=0}^{\infty} 2^{2m\alpha} E_{2^m}(f)_{\mathcal{H}}^2 \le \sum_{m=0}^{\infty} 2^{2m\alpha} \sum_{k=m+1}^{\infty} F_k^2 \le C \sum_{k=1}^{\infty} 2^{2k\alpha} F_k^2,$$

which gives the other implication of the asserted equivalence.

Let us digest these results with the following example. We take for $\mathcal{H}$ the space $L_2(\mathbb{T})$ of 2π-periodic functions on the unit circle $\mathbb{T}$ which has the Fourier basis $\{(2\pi)^{-\frac{1}{2}}e^{ikx} : k \in \mathbb{Z}\}$. (Note here the indexing of the basis functions on $\mathbb{Z}$ rather than $\mathbb{N}$.) The space $\mathcal{H}_n := \operatorname{span}\{e^{ikx} : |k| \le n\}$ is the space $\mathcal{T}_n$ of trigonometric polynomials of degree $\le n$. The coefficients with respect to this basis are the Fourier coefficients $\hat{f}(k)$ and therefore (2.13) states that $\mathcal{A}_2^\alpha((\mathcal{T}_n))$ is characterized by the condition

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} |k|^{2\alpha}|\hat{f}(k)|^2 \le M. \tag{2.15}$$

If α is an integer, (2.15) describes the Sobolev space $W^\alpha(L_2(\mathbb{T}))$ of all 2π-periodic function with their αth derivative in $L_2(\mathbb{T})$ and the sum in (2.15) is the square of the semi-norm $|f|_{W_2^\alpha(L_2(\mathbb{T}))}$. For noninteger α, (2.15) characterizes, by definition, the fractional order Sobolev space $W^\alpha(L_2(\mathbb{T}))$. One should note that one half of the characterization (2.15) of $\mathcal{A}_2^\alpha((\mathcal{T}_n))$ gives the inequality

$$\left(\sum_{n=1}^{\infty}[n^\alpha E_n(f)_{\mathcal{H}}]^2\frac{1}{n}\right)^{1/2} \le C|f|_{W^\alpha(L_2(\mathbb{T}))} \tag{2.16}$$

which is slightly stronger than the inequality

$$E_n(f)_{\mathcal{H}} \le Cn^{-\alpha}|f|_{W^\alpha(L_2(\mathbb{T}))}, \tag{2.17}$$

which is more frequently found in the literature.

Using (2.10), it is easy to prove that the space $\mathcal{A}^\alpha((\mathcal{T}_n))$ is identical with the Besov space $B_\infty^\alpha(L_2(\mathbb{T}))$ and, for noninteger α, this is the Lipschitz space $\operatorname{Lip}(\alpha, L_2(\mathbb{T}))$. (We introduce and discuss amply the Besov and Lipschitz spaces in Sections 3.2 and 4.5.)

Let us return now to the case of a general Hilbert space $\mathcal{H}$ and nonlinear approximation from Σ_n. We can characterize the space $\mathcal{A}^\alpha((\Sigma_n))$ by using the rearrangement of the coefficients f_k. We denote by $\gamma_k(f)$ the kth largest of the numbers $|f_j|$. We first want to observe that $f \in \mathcal{A}^\alpha((\Sigma_n))$ if and only if

$$\gamma_n(f) \le Mn^{-\alpha-1/2} \tag{2.18}$$

and the infimum of all M which satisfy (2.18) is equivalent to $|f|_{\mathcal{A}^\alpha((\Sigma_n))}$. Indeed, we have

$$\sigma_n(f)_{\mathcal{H}}^2 = \sum_{k>n}\gamma_k(f)^2. \tag{2.19}$$

Therefore, if f satisfies (2.18), then clearly

$$\sigma_n(f)_{\mathcal{H}} \le CMn^{-\alpha},$$

so that $f \in \mathcal{A}^\alpha((\Sigma_n))$ and we have one of the implications in the asserted characterization. On the other hand, if $f \in A^\alpha((\Sigma_n))$, then

$$\gamma_{2n}(f)^2 \le n^{-1} \sum_{m=n+1}^{2n} \gamma_m(f)^2 \le n^{-1}\sigma_n(f)_{\mathcal{H}}^2 \le |f|^2_{\mathcal{A}^\alpha((\Sigma_n))} n^{-2\alpha-1}.$$

Since a similar inequality holds for $\gamma_{2n+1}(f)$, we have the other implication of the asserted equivalence.

It is also easy to characterize other approximation classes such as the $\mathcal{A}_2^\alpha((\Sigma_n))$, which is the analogue of $\mathcal{A}_2^\alpha((\mathcal{H}_n))$. We shall formulate such results in Section 5.

Let us return to our example of trigonometric approximation. Approximation by Σ_n is n-term approximation by trigonometric sums. It is easy to see the distinction between linear and nonlinear approximation in this case. Linear approximation corresponds to a certain decay in the Fourier coefficients $\hat{f}(k)$ as the frequency k increases, whereas nonlinear approximation corresponds to a decay in the rearranged coefficients. Thus, nonlinear approximation does not recognize the frequency location of the coefficients. If we reassign the Fourier coefficients of a function $f \in \mathcal{A}^\alpha$ to new frequency locations, the resulting function is still in $\mathcal{A}^\alpha$. Thus, in the nonlinear case there is no correspondence between rate of approximation to classical smoothness as there was in the linear case. It is possible to have large coefficients at high frequency just as long as there are not too many of them. For example, the functions e^{ikx} are obviously in all of the spaces $\mathcal{A}^\alpha$ even though their derivatives are large when k is large.

3. Approximation by piecewise constants

For our next taste of nonlinear approximation, we shall consider in this section several types of approximation by piecewise constants corresponding to linear and nonlinear approximation. Our goal is to see in this very simple setting the advantages of nonlinear methods. We begin with a target function f defined on $\Omega := [0, 1)$ and approximate it in various ways by piecewise constants with n pieces. We shall be interested in the efficiency of such approximation, that is, how the error of approximation decreases as n tends to infinity. We shall see that, in many cases, we can characterize the functions f which have certain approximation orders (for instance $O(n^{-\alpha})$, $0 < \alpha \le 1$). Such characterizations will illuminate the distinctions between linear and nonlinear approximation.

3.1. Linear approximation by piecewise constants

We begin by considering approximation by piecewise constants on partitions of Ω which are fixed in advance. This will be our reference point for comparisons with nonlinear approximations that follow. This form of linear approximation is also important in numerical computation since it is the simplest setting for FEM and other numerical methods based on approximation by piecewise polynomials. We shall see that there is a complete understanding in this case of the properties of the target function needed to guarantee certain approximation rates. As we shall amplify below, this theory explains what we should be able to achieve with proper numerical methods and also tells us what form good numerical estimates should take.

Let N be a positive integer and let $T := \{0 =: t_0 < t_1 < \cdots < t_N := 1\}$ be an ordered set of points in Ω. These points determine a partition $\Pi := \Pi(T) := \{I_k\}_{k=1}^N$ of Ω into N disjoint intervals $I_k := [t_{k-1}, t_k)$, $1 \le k \le N$. Let $\mathcal{S}^1(T)$ denote the space of piecewise constant functions relative to this partition. The characteristic functions $\{\chi_I : I \in \Pi\}$ form a basis for $\mathcal{S}^1(T)$: each function $S \in \mathcal{S}^1(T)$ can be represented uniquely by

$$S = \sum_{I \in \Pi} c_I \chi_I. \tag{3.1}$$

Thus $\mathcal{S}^1(T)$ is a linear space of dimension N.

For $0 < p \le \infty$, we introduce the error in approximating a function $f \in L_p[0,1)$ by the elements of $\mathcal{S}^1(T)$:

$$s(f,T)_p := \inf_{S \in \mathcal{S}^1(T)} \|f - S\|_{L_p[0,1)}. \tag{3.2}$$

We would like to understand what properties of f and T determine $s(f,T)_p$. For the moment, we shall restrict our discussion to the case $p = \infty$ which corresponds to uniformly continuous functions f on $[0,1)$ to be approximated in the uniform norm (L_∞-norm) on $[0,1)$. The quality of approximation that $S'(T)$ provides is related to the mesh length

$$\delta_T := \max_{0 \le k < N} |t_{k+1} - t_k|. \tag{3.3}$$

We shall first give estimates for $s(f,T)_\infty$ and then later ask in what sense these estimates are best possible. We recall the definition of the Lipschitz spaces $\mathrm{Lip}\,\alpha$. For each $0 \le \alpha \le 1$ and $M > 0$, we let $\mathrm{Lip}_M\,\alpha$ denote the set of all functions f on Ω such that

$$|f(x) - f(y)| \le M|x - y|^\alpha.$$

Then $\mathrm{Lip}\,\alpha := \cup_{M>0} \mathrm{Lip}_M\,\alpha$. The infimum of all M for which $f \in \mathrm{Lip}_M\,\alpha$ is by definition $|f|_{\mathrm{Lip}\,\alpha}$. In particular, $f \in \mathrm{Lip}\,1$ if and only if f is absolutely continuous and $f' \in L_\infty$; moreover, $|f|_{\mathrm{Lip}\,1} = \|f'\|_{L_\infty}$.

If the target function $f \in \mathrm{Lip}_M\,\alpha$, then

$$s(f,T)_\infty \le M(\delta_T/2)^\alpha. \tag{3.4}$$

Indeed, we define the piecewise constant function $S \in \mathcal{S}^1(T)$ by

$$S(x) := f(\xi_I), \quad x \in I,\ I \in \Pi_n,$$

with ξ_I the midpoint of I. Then, $|x-\xi_I| \le \delta_T/2$, $x \in I$, and hence

$$\|f-S\|_{L_\infty[0,1)} \le M(\delta_T/2)^\alpha, \tag{3.5}$$

which gives (3.4).

We turn now to the question of whether the estimate (3.4) is the best we can do. We shall see that this is indeed the case in several senses. First, suppose that for a function f we know that

$$s(f,T)_\infty \le M\delta_T^\alpha, \tag{3.6}$$

for *every* partition T. Then, we can prove that f is in $\mathrm{Lip}\,\alpha$ and moreover $|f|_{\mathrm{Lip}\,\alpha} \le M$. Results of this type are called inverse theorems in approximation theory whereas results like (3.4) are called direct theorems.

To prove the inverse theorem, we need to estimate the smoothness of f from the approximation errors $s(f,T)_\infty$. In the case at hand, the proof is very simple. Let S_T be a best approximation to f from $\mathcal{S}^1(T)$ in the $L_\infty(\Omega)$-norm. (A simple compactness argument shows the existence of best approximants.) If x, y are two points from Ω that are in the same interval $I \in \Pi(T)$, then from (3.6)

$$\begin{aligned} |f(x)-f(y)| \;&\le\; |f(x)-S_T(x)| + |f(y)-S_T(y)| \\ &+\; |S_T(x)-S_T(y)| \le 2s(f,T)_\infty \le 2M\delta_T^\alpha \end{aligned} \tag{3.7}$$

because $S_T(x) = S_T(y)$ (S_T is constant on I). Since we can choose T so that δ_T is arbitrarily close to $|x-y|$, we obtain

$$|f(x)-f(y)| \le 2M(\delta_T)^\alpha \le 2M|x-y|^\alpha \tag{3.8}$$

which shows that $f \in \mathrm{Lip}\,\alpha$ and $|f|_{\mathrm{Lip}\,\alpha} \le 2M$.

Here is one further observation on the above analysis. If f is a function for which $s(f,T)_\infty = o(\delta_T)$ holds for all T, then the above argument gives that $f(x+h)-f(x) = o(h)$, $h \to 0$, for each $x \in \Omega$. Thus f is constant (its derivative is 0 everywhere). This is called a saturation theorem in approximation theory. Only trivial functions can be approximated with order better than $O(\delta_T)$.

The above discussion is not completely satisfactory for numerical analysis. In numerical algorithms, we usually have only a sequence of partitions. However, with some massaging, the above arguments can be applied in this case as well. Consider, for example, the case where

$$\Delta_n := \{k/n : 0 \le k \le n\} \tag{3.9}$$

consists of n equally spaced points from Ω (with spacing $1/n$). Then, for each $0 < \alpha \le 1$, a function f satisfies

$$s_n(f)_\infty := s(f,\Delta_n)_\infty = O(n^{-\alpha}) \tag{3.10}$$

if and only if $f \in \mathrm{Lip}\,\alpha$ (see DeVore and Lorentz (1993)). The saturation result holds as well. If $s_n(f)_\infty = o(n^{-1})$ then f is constant. Of course the direct estimates in this setting follow from (3.4). The inverse estimates are a little more subtle and use the fact that the sets Δ_n mix; that is, each point $x \in (0,1)$ falls in the 'middle' of many intervals from the partitions associated to Δ_n. If we consider partitions that do not mix then, while direct estimates are equally valid, the inverse estimates generally fail. A case in point are the dyadic partitions whose sets of breakpoints Δ_{2^n} are nested. A piecewise constant function from $S^1(\Delta_{2^n})$ will be approximated exactly by elements from $S^1(\Delta_{2^m})$, $m \geq n$, and yet these functions are not even continuous.

An analysis similar to that given above holds for approximation in L_p, for $1 \leq p < \infty$, and even for $0 < p < 1$. To explain these results, we define the space $\mathrm{Lip}(\alpha, L_p(\Omega))$, $0 < \alpha \leq 1$, $0 < p \leq \infty$, which is the set of all functions $f \in L_p(\Omega)$ for which

$$\|f(\cdot + h) - f\|_{L_p[0,1-h)} \leq M h^\alpha, \quad 0 < h < 1. \tag{3.11}$$

Again, the smallest $M \geq 0$ for which (3.11) holds is $|f|_{\mathrm{Lip}(\alpha, L_p(\Omega))}$.

By analogy with (3.4), there are $S_T \in \mathcal{S}^1(T)$ such that

$$s(f,T)_p \leq \|f - S_T\|_{L_p(\Omega)} \leq C_p |f|_{\mathrm{Lip}(\alpha, L_p(\Omega))} \delta_T^\alpha \tag{3.12}$$

with the constant C_p depending at most on p. Indeed, for $p \geq 1$, we can define S_T by

$$S_T(x) := a_I(f), \quad x \in I, \ I \in \Pi(T), \tag{3.13}$$

with[2]

$$a_I(f) := \frac{1}{|I|} \int_I f \, \mathrm{d}x$$

the average of f over I. With this definition of S_T one easily derives (3.12); see Section 2 of Chapter 12 in DeVore and Lorentz (1993). When $0 < p < 1$, we replace $a_I(f)$ by the median of f on the interval I (see Brown and Lucier (1994)).

Inverse estimates follow the same lines as the case $p = \infty$ discussed above. We limit further discussion to the case Δ_n of equally spaced breakpoints given by (3.9). Then, if f satisfies

$$s_n(f)_p := s(f, \Delta_n)_p \leq M n^{-\alpha}, \quad n = 1, 2, \ldots, \tag{3.14}$$

for some $0 < \alpha \leq 1$, $M > 0$, then $f \in \mathrm{Lip}(\alpha, L_p(\Omega))$ and

$$|f|_{\mathrm{Lip}(\alpha, L_p(\Omega))} \leq C_p M.$$

[2] We shall use the notation $|E|$ to denote the Lebesgue measure of a set E throughout this paper.

The saturation theorem is also valid: if $s_n(f)_p = o(n^{-1})$, $n \to \infty$, then f is constant.

In summary, we know precisely when a function satisfies $s_n(f)_p = O(n^{-\alpha})$, $n = 1, 2, \ldots$; it should be in the space $\mathrm{Lip}(\alpha, L_p(\Omega))$. This provides a guide to the construction and analysis of numerical methods based on approximation by piecewise constants. For example, suppose that we are using $\mathcal{S}^1(\Delta_n)$ to generate a numerical approximation $A_n u$ to a function u which is known to be in $\mathrm{Lip}(1, L_p(\Omega))$. The values of u would not be known to us but would be generated by our numerical method. The estimates (3.4) or (3.12) tell us what we could expect of the numerical method in the best of all worlds. If we are able to prove that our numerical method satisfies

$$\|u - A_n u\|_{L_p(\Omega)} \le C_p |f|_{\mathrm{Lip}(1, L_p(\Omega))} n^{-1}, \quad n = 1, 2, \ldots, \tag{3.15}$$

we can rest assured that we have done the best possible (save for the numerical constant C_p). If we cannot prove such an estimate then we should try to understand why. Moreover, (3.15) is the correct form of error estimates based on approximation by piecewise constants on uniform partitions.

There are numerous generalizations of the results given in this section. First of all, piecewise constants can be replaced by piecewise polynomials of degree r with r arbitrary but fixed (see Section 6.2). One can require that the piecewise polynomials have smoothness C^{r-2} at the breakpoints with an identical theory. Of course, inverse theorems still require some mixing condition. Moreover, all of these results hold in the multivariate case as is discussed in Section 6.2. We can also do a more subtle analysis of approximation orders where $O(n^{-\alpha})$ is replaced by a more general statement on the rate of decay of the error. This is important for a fuller understanding of approximation theory and its relationship to function spaces. We shall discuss these issues in Section 4 after the reader has more familiarity with more fundamental approximation concepts.

3.2. Nonlinear approximation by piecewise constants

In linear approximation by piecewise constants, the partitions are chosen in advance and are independent of the target function f. The question arises whether there is anything to be gained by allowing the partition to depend on f. This brings us to try to understand approximation by piecewise constants where the number of pieces is fixed but the actual partition can vary with the target function. This is the simplest case of what is called variable knot spline approximation. It is also one of the simplest and most instructive examples of nonlinear approximation.

If T is a finite set of points $0 =: t_0 < t_1 < \cdots < t_n := 1$ from Ω, we denote by $\mathcal{S}^1(T)$ the functions S which are piecewise constant with breakpoints from T. Let $\Sigma_n := \cup_{\#T=n+1} \mathcal{S}^1(T)$, where $\#T$ denotes the cardinality of T.

Each function in Σ_n is piecewise constant with at most n pieces. Note that Σ_n is not a linear space; for example, adding two functions in Σ_n results in a piecewise constant function but with as many as $2n$ pieces. Given $f \in L_p(\Omega)$, $0 < p \le \infty$, we introduce

$$\sigma_n(f)_p := \inf_{S \in \Sigma_n} \|f - S\|_{L_p(\Omega)}, \tag{3.16}$$

which is the L_p-error of nonlinear piecewise constant approximation to f.

As noted earlier, we would like to understand what properties of f determine the rate of decrease of $\sigma_n(f)_p$. We shall begin our discussion with the case $p = \infty$, which corresponds to approximating the continuous function f in the uniform norm. We shall show the following result of Kahane (1961). For a function $f \in C(\Omega)$ we have

$$\sigma_n(f)_\infty \le \tfrac{M}{2n}, \quad n = 1, 2, \ldots, \tag{3.17}$$

if and only if $f \in \mathrm{BV}$, *i.e.*, f, is of bounded variation on Ω and $|f|_{\mathrm{BV}} := \mathrm{Var}_\Omega(f)$ is identical with the smallest constant M for which (3.17) holds.

We sketch the proof of Kahane's result since it is quite simple and instructive. Suppose first that $f \in \mathrm{BV}$ with $M := \mathrm{Var}_\Omega(f)$. Since f is, by assumption, continuous, we can find $T := \{0 =: t_0, \ldots, t_n := 1\}$ such that $\mathrm{Var}_{[t_{k-1}, t_k)} f \le M/n$, $k = 1, \ldots, n$. If a_k is the median value of f on $[t_{k-1}, t_k]$, and $S_n(x) := a_k$, $x \in [t_{k-1}, t_k)$, $k = 1, \ldots, n$, then $S_n \in \Sigma_n$ and satisfies

$$\|f - S_n\|_{L_\infty(\Omega)} \le M/2n, \tag{3.18}$$

which shows (3.17).

Conversely, suppose that (3.17) holds for some $M > 0$. Let $S_n \in \Sigma_n$ satisfy $\|f - S_n\|_{L_\infty(\Omega)} \le (M + \epsilon)/(2n)$ with $\epsilon > 0$. If $x_0 := 0 < x_1 < \cdots < x_m := 1$ is an arbitrary partion for Ω and ν_k is the number of values that S_n attains on $[x_{k-1}, x_k)$, then one easily sees that

$$|f(x_k) - f(x_{k-1})| \le 2\nu_k \|f - S_n\|_{L_\infty(\Omega)} \le \tfrac{\nu_k(M+\epsilon)}{n}, \quad k = 1, 2, \ldots, m. \tag{3.19}$$

Since $\sum_{k=1}^m \nu_k \le m + n$, we have

$$\sum_{k=1}^m |f(x_k) - f(x_{k-1})| \le \sum_{k=1}^m \tfrac{\nu_k(M+\epsilon)}{n} \le (M + \epsilon)(1 + \tfrac{m}{n}). \tag{3.20}$$

Letting $n \to \infty$ and then $\epsilon \to 0$ we find

$$\sum_{k=1}^m |f(x_k) - f(x_{k-1})| \le M, \tag{3.21}$$

which shows that $\mathrm{Var}_\Omega(f) \le M$.

There are elements of the above proof that are characteristic of nonlinear approximation. Firstly, the partition providing (3.17) depends on f.

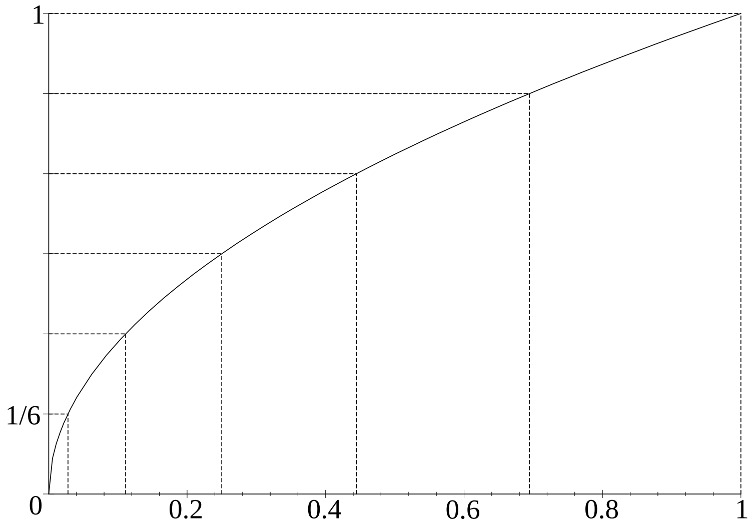

Fig. 1. Best selection of breakpoints for $f(x) = x^{1/2}$ when $n = 6$

Secondly, this partition is obtained by balancing the variation of f over the intervals I in this partition. In other types of nonlinear approximation, $\mathrm{Var}_I(f)$ will be replaced by some other expression $B(f, I)$ defined on intervals I (or other sets in more general settings).

Let us pause now for a moment to compare Kahane's result with what we know about linear approximation by piecewise constants in the uniform norm. In both cases, we can characterize functions which can be approximated with efficiency $O(n^{-1})$. In the case of linear approximation from $\mathcal{S}^1(T_n)$ (as described in the previous section), this is the class of functions $\mathrm{Lip}\,(1, L_\infty(\Omega))$ or, equivalently, functions f for which $f' \in L_\infty(\Omega)$. On the other hand, for nonlinear approximation, it is the class BV of functions of bounded variation. It is well known that $\mathrm{BV} = \mathrm{Lip}(1, L_1(\Omega))$ with equivalent norms. Thus in both cases the function is required to have one order of smoothness but measured in quite different norms. For linear approximation the smoothness is measured in L_∞, the same norm as the underlying approximation. For nonlinear approximation the smoothness is measured in L_1. Thus, in nonlinear approximation, the smoothness is measured in a weaker norm. What is the significance of L_1? The answer lies in the Sobolev embedding theorem. Among the spaces $\mathrm{Lip}(1, L_p(\Omega))$, $0 < p \le \infty$, $p = 1$ is the smallest value for which this space is embedded in $L_\infty(\Omega)$. In other words, the functions in $\mathrm{Lip}(1, L_1(\Omega))$ barely get into $L_\infty(\Omega)$ (the space in which we measure error) and yet we can approximate them quite well.

An example might be instructive. Consider the function $f(x) = x^\alpha$ with $0 < \alpha < 1$. This function is in $\mathrm{Lip}(\alpha, L_\infty(\Omega))$ and in no higher-order Lipschitz space. It can be approximated by elements of $\mathcal{S}^1(T_n)$ with order

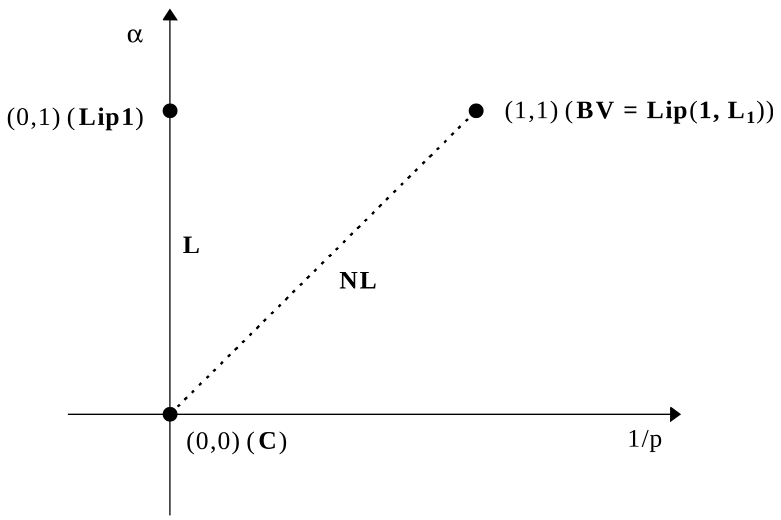

Fig. 2. Linear and nonlinear approximation in C

exactly $O(n^{-\alpha})$. On the other hand, this function is clearly of bounded variation (being monotone) and hence can be approximated by the elements of Σ_n to order $O(n^{-1})$. It is easy to see how to construct such an approximant. Consider the graph of f as depicted in Figure 1. We divide the range of f (which is the interval $[0, 1)$) on the y-axis into n pieces corresponding to the y values $y_k := k/n$, $k = 0, 1, \ldots, n$. The preimage of these points is the set $\{x_k := (k/n)^{1/\alpha} : 0 \le k \le n\}$, which forms our set T of breakpoints for the best piecewise polynomial approximant from Σ_n.

It will be useful to have a way of visualizing spaces of functions as they occur in our discussion of approximation. This will give us a simple way to keep track of various results and also add to our understanding. We shall do this by using points in the upper right quadrant of the plane. The x-axis will correspond to the L_p spaces except that L_p is identified with $x = 1/p$ not with $x = p$. The y axis will correspond to the order of smoothness. For example $y = 1$ will mean a space of smoothness order one (or one time differentiable, if you like). Thus $(1/p, \alpha)$ corresponds to a space of smoothness α measured in the L_p-norm. For example, we could identify this point with the space $\mathrm{Lip}(\alpha, L_p)$ although when we get to finer aspects of approximation theory we may want to vary this interpretation slightly.

Figure 2 gives a summary of our knowledge so far. The vertical line segment (marked L) connecting $(0, 0)$ (L_∞) to $(0, 1)$ $(\mathrm{Lip}(1, L_\infty))$ correspond to the spaces we engaged when we characterized approximation order for linear approximation (approximation from $\mathcal{S}^1(T_n)$). For example, $(0, 1)$ $(\mathrm{Lip}(1, L_\infty))$ was the space of functions with approximation order $O(n^{-1})$.

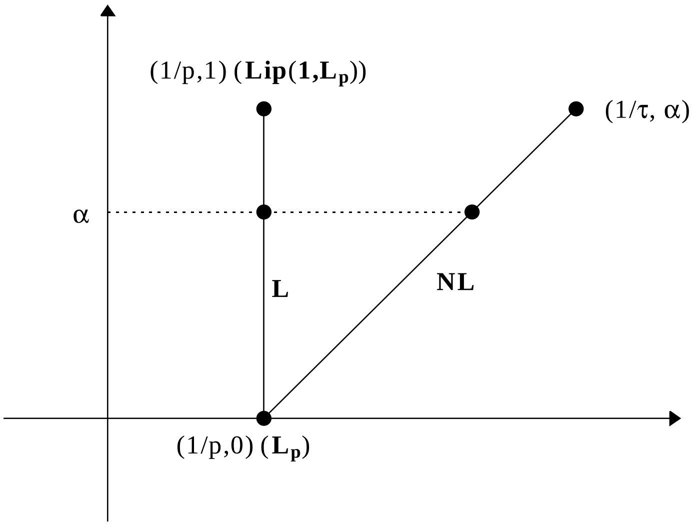

Fig. 3. Linear and nonlinear approximation in L_p

On the other hand, for nonlinear approximation from Σ_n, we saw that the point $(1,1)$ ($\mathrm{Lip}(1,L_1)$) describes the space of functions which are approximated with order $O(n^{-1})$. We shall see later (Section 4) that the point (α,α) on the line connecting $(0,0)$ to $(1,1)$ (marked NL) describes the space of functions approximated with order $O(n^{-\alpha})$ (a few new wrinkles come in here which is why we are postponing a precise discussion).

More generally, approximation in L_p, $0 < p \le \infty$, is depicted in Figure 3. The spaces corresponding to linear approximation lie on the vertical line segment (marked L) connecting $(1/p,0)$ (L_p) to $(1/p,1)$ ($\mathrm{Lip}(1,L_p)$), whereas the line segment (marked NL) emanating from $(1/p,0)$ with slope one will describe the nonlinear approximation spaces. The points on this line are of the form $(1/\tau,\alpha)$ with $1/\tau = \alpha + 1/p$. Again, this line segment in nonlinear approximation corresponds to the limiting spaces in the Sobolev embedding theorem. Spaces to the left of this line segment are embedded into L_p; those to the right are not.

There are various generalizations of nonlinear piecewise constant approximation which we shall address in due course. For univariate approximation, we can replace piecewise constant functions by piecewise polynomials of fixed degree r with n free knots with a similar theory (Section 6.3). However, multivariate approximation by piecewise polynomials leads to new difficulties, as we shall see in Section 6.5.

Approximation by piecewise constants (or more generally piecewise polynomials) with free knots is used in numerical PDEs. It is particularly useful when the solution is known to develop singularities. An example would be a nonlinear transport equation in which shocks appear (see Section 10). The significance of the above results to the numerical analyst is that it clarifies what is the optimal performance that can be obtained by such methods. Once the norm has been chosen in which the error is to be measured, then we understand the minimal smoothness that will allow a given approximation rate. We also understand what form error estimates should take. For example, consider numerically approximating a function u by a piecewise constant function $A_n u$ with n free knots. We have seen that, in the case of uniform approximation, the correct form of the error estimate is

$$\|u - A_n u\|_{L_\infty(\Omega)} \leq C \frac{|u|_{\mathrm{BV}}}{n}. \tag{3.22}$$

This is in contrast to the case of fixed knots where $|u|_{\mathrm{BV}}$ is replaced by $\|u'\|_{L_\infty(\Omega)}$. A similar situation exists when error is measured in other L_p-norms, as will be developed in Section 6.

The above theory of nonlinear piecewise constant approximation also tells us the correct form for local error estimators. Approximating in L_∞, we should estimate local error by local variation. Approximating in L_p, the variation will be replaced by other set functions obtained from certain Besov or Sobolev norms (see Section 6.1).

3.3. Adaptive approximation by piecewise constants

One disadvantage of piecewise constant approximation with free knots is that it is not always easy to find partitions that realize the optimal approximation order. This is particularly true in the case of numerical approximation when the target function is not known to us but is only approximated as we proceed numerically. One way to ameliorate this situation is to generate partitions adaptively. New breakpoints are added as new information is gained about the target function. We shall discuss this type of approximation in this section with the goal of understanding what is lost in terms of accuracy of approximation when adaptive partitions are used in place of free partitions. Adaptive approximation is also important because it generalizes readily to the multivariate case when intervals are replaced by cubes.

The starting point for adaptive approximation is a function $\mathcal{E}(I)$ which is defined for each interval $I \subset \Omega$ and estimates the approximation error on I. Namely, let $E(f, I)_p$ be the local error in approximating f by constants in the $L_p(I)$-norm:

$$E(f, I)_p := \inf_{c \in \mathbb{R}} \|f - c\|_{L_p(I)}. \tag{3.23}$$

Then, we assume that $\mathcal{E}$ satisfies

$$E(f, I)_p \leq \mathcal{E}(I). \tag{3.24}$$

In numerical settings, $\mathcal{E}(I)$ is an upper bound for $E(f,I)_p$ obtained from the information at hand. It is at this point that approximation theory and numerical analysis sometimes part company. Approximation theory assumes enough about the target function to have an effective error estimator $\mathcal{E}$, a property not always verifiable for numerical estimators.

To retain the spirit of our previous sections, let us assume for our illustration that $p=\infty$ so that we are approximating continuous functions in the $L_\infty(\Omega)$ norm. In this case, a simple upper bound for $E(f,I)_\infty$ is provided by

$$E(f,I)_\infty \le \mathrm{Var}_I(f) \le \int_I |f'(x)|\,\mathrm{d}x, \tag{3.25}$$

which holds whenever these quantities are defined for the continuous function f (*i.e.*, f should be in BV for the first estimate, $f' \in L_1$ for the second). Thus, we could take for $\mathcal{E}$ any of the three quantities appearing in (3.25). A common feature of each of these error estimators is that

$$\mathcal{E}(I_1)+\mathcal{E}(I_2) \le \mathcal{E}(I_1 \cup I_2), \quad I_1 \cap I_2 = \emptyset. \tag{3.26}$$

We shall restrict our attention to adaptive algorithms that create partitions of Ω consisting of dyadic intervals. Our development parallels completely the standard treatment of adaptive numerical quadrature. We shall denote by $D := D(\Omega)$ the set of all dyadic intervals in Ω; for specificity we take these intervals to be closed on the left end-point and open on the right. Each interval $I \in D$ has two *children*. These are the intervals $J \in D$ such that $J \subset I$ and $|J| = |I|/2$. If J is a child of I then I is called the *parent* of J. Intervals $J \in D$ such that $J \subset I$ are *descendants* of I those with $I \subset J$ are *ancestors* of I.

A typical adaptive algorithm proceeds as follows. We begin with our target function f, an error estimator $\mathcal{E}$, and a target tolerance ϵ which relates to the final approximation error we want to attain. At each step of the algorithm we have a set $\mathcal{G}$ of good intervals (on which the local error meets the tolerance) and a set $\mathcal{B}$ of bad intervals (on which we do not meet the tolerance). Good intervals become members of our final partition. Bad intervals are further processed: they are halved and their children are checked for being good or bad.

Initially, we check $\mathcal{E}(\Omega)$. If $\mathcal{E}(\Omega) \le \epsilon$ then we define $\mathcal{G} = \{\Omega\}$, $\mathcal{B} := \emptyset$ and we terminate the algorithm. On the other hand, if $\mathcal{E}(\Omega) > \epsilon$, we define $\mathcal{G} = \emptyset$, $\mathcal{B} := \{\Omega\}$ and proceed with the following general step of the algorithm.

General step. Given any interval I in the current set $\mathcal{B}$ of bad intervals, we process it as follows. For each of the two children J of I, we check $\mathcal{E}(J)$. If $\mathcal{E}(J) \le \epsilon$, then J is added to the set of good intervals. If $E(J) > \epsilon$, then J is added to the set of bad intervals. Once a bad interval is processed it is removed from $\mathcal{B}$.

The algorithm terminates when $\mathcal{B} = \emptyset$, and the final set of good intervals is denoted by $\mathcal{G}_\epsilon := \mathcal{G}_\epsilon(f)$. The intervals in $\mathcal{G}_\epsilon$ form a partition of Ω, that is, they are pairwise disjoint and their union is all of Ω. We define

$$S_\epsilon := \sum_{I \in \mathcal{G}_\epsilon} c_I \chi_I, \tag{3.27}$$

where c_I is a constant that satisfies

$$\|f - c_I\|_{L_\infty(I)} \le \mathcal{E}(I) \le \epsilon, \quad I \in \mathcal{G}_\epsilon.$$

Thus, S_ϵ is a piecewise constant function approximating f to tolerance ϵ:

$$\|f - S_\epsilon\|_{L_\infty(\Omega)} \le \epsilon. \tag{3.28}$$

The approximation efficiency of the adaptive algorithm depends on the number $N_\epsilon(f) := \#\mathcal{G}_\epsilon(f)$ of good intervals. We are interested in estimating N_ϵ so that we can compare adaptive efficiency with free knot spline approximation. For this we recall the space $L \log L$, which consists of all integrable functions for which

$$\|f\|_{L \log L} := \int_\Omega |f(x)|(1 + \log |f(x)|)\,\mathrm{d}x$$

is finite. This space contains all spaces L_p, $p > 1$, but is strictly contained in $L_1(\Omega)$. We have shown in DeVore (1987) that any of the three estimators of (3.25) satisfy

$$N_\epsilon(f) \le C \frac{\|f'\|_{L \log L}}{\epsilon}. \tag{3.29}$$

We shall give the proof of (3.29), which is not difficult. It will allow us to introduce some concepts that are useful in nonlinear approximation and numerical estimation, such as the use of maximal functions. The Hardy–Littlewood maximal function Mf is defined for a function in $L_1(\Omega)$ by

$$Mf(x) := \sup_{I \ni x} \tfrac{1}{|I|} \int_I |f(y)|\,\mathrm{d}y, \tag{3.30}$$

where the sup is taken over all intervals $I \subset \Omega$ which contain x. Thus $Mf(x)$ is the smallest number that bounds all of the averages of $|f|$ over intervals which contain x. The maximal function Mf is at the heart of differentiability of functions (see Chapter 1 of Stein (1970)). We shall need the fact (see pages 243–246 of Bennett and Sharpley (1988)) that

$$\|f\|_{L \log L} \asymp \int_\Omega Mf(y)\,\mathrm{d}y. \tag{3.31}$$

We shall use Mf to count N_ϵ. We assume that $\mathcal{G}_\epsilon \neq \{\Omega\}$. Suppose that $I \in \mathcal{G}_\epsilon$. Then the parent J of I satisfies

$$\epsilon < \mathcal{E}(J) \leq \int_J |f'(y)|\,\mathrm{d}y \leq |J| Mf'(x), \tag{3.32}$$

for all $x \in J$. In particular, we have

$$\epsilon \leq |J| \inf_{x\in I} Mf'(x) \leq \frac{|J|}{|I|}\int_I Mf'(y)\,\mathrm{d}y = 2\int_I Mf'(y)\,\mathrm{d}y. \tag{3.33}$$

Since the intervals in $\mathcal{G}_\epsilon$ are disjoint, we have

$$N_\epsilon \epsilon \leq 2\sum_{I\in\mathcal{G}_\epsilon}\int_I Mf'(y)\,\mathrm{d}y = 2\int_\Omega Mf'(y)\,\mathrm{d}y \leq C\|f'\|_{L\log L},$$

where the last inequality uses (3.31). This proves (3.29).

In order to compare adaptive approximation with free knot splines, we introduce the adaptive approximation error

$$a_n(f)_\infty := \inf\{\epsilon : N_\epsilon(f) \leq n\}. \tag{3.34}$$

Thus, with the choice $\epsilon = (C\|f'\|_{L\log L})/n$, and C the constant in (3.29), our adaptive algorithm generates a partition $\mathcal{G}$ with at most n dyadic intervals and, from (3.28), we have

$$a_n(f)_\infty \leq \|f - S_\epsilon\|_{L_\infty(\Omega)} \leq C\frac{\|f'\|_{L\log L}}{n}. \tag{3.35}$$

Let's compare $a_n(f)_\infty$ with the error $\sigma_n(f)_\infty$ for free knot approximation. In free knot splines we obtained the approximation rate $\sigma_n(f)_\infty = O(n^{-1})$ if and only if $f \in \mathrm{BV}$. This condition is slightly weaker than requiring that f' is in $L_1(\Omega)$ (the derivative of f should be a Borel measure). On the other hand, assuming that f satisfies the slightly stronger condition $f' \in L\log L$, we find $a_n(f)_\infty \leq C/n$. Thus, the cost in using adaptive algorithms is slight from the viewpoint of the smoothness condition required on f to produce the order $O(n^{-1})$.

It is much more difficult to prove error estimates for numerically based adaptive algorithms. What is needed is a comparison (from above and below) of the error estimator $\mathcal{E}(I)$ with the local approximation error $E(f,I)_p$ or one of the good estimators like $\int_I |f'|$. Nevertheless, the above results are useful in that they give the form such error estimators $\mathcal{E}(I)$ should take and also give the form the error analysis should take.

There is a comparable theory for adaptive approximation in other L_p-norms and even in several variables (Birman and Solomyak 1967).

3.4. n-term approximation: a first look

There is another view toward the results we have obtained thus far, which is important because it generalizes readily to a variety of settings. In each of the three types of approximation (linear, free knot, and adaptive), we have constructed an approximant of the form

$$S = \sum_{I\in\Lambda} c_I \chi_I, \tag{3.36}$$

where Λ is a set of intervals and the c_I are constants. Thus, a general approximation problem that would encompass all three of the above is to approximate using sums (3.36) where $\#\Lambda \le n$. This is called *n-term approximation*. We formulate this problem more formally as follows.

Let Σ_n^* be the set of all piecewise constant functions that can be written as in (3.36) with $\#\Lambda \le n$. Then, Σ_n^* is a nonlinear space. As in our previous considerations, we define the L_p-approximation error

$$\sigma_n^*(f)_p := \inf_{S\in\Sigma_n^*} \|f - S\|_{L_p(\Omega)}. \tag{3.37}$$

Note that we do not require that the intervals of Λ form a disjoint partition; we allow possible overlap in the intervals.

It is easy to see that the approximation properties of n-term approximation is equivalent to that of free knot approximation. Indeed, $\Sigma_n \subset \Sigma_n^* \subset \Sigma_{2n}$, $n = 1, 2, \ldots$, and therefore

$$\sigma_{2n}(f)_p \le \sigma_n^*(f)_p \le \sigma_n(f)_p. \tag{3.38}$$

Thus, for example, a function f satisfies $\sigma_n^*(f)_p = O(n^{-\alpha})$ if and only if $\sigma_n^*(f)_p = O(n^{-\alpha})$.

The situation with adaptive algorithms is more interesting and enlightening. In analogy to the above, one defines Σ_n^a as the set of functions S which can be expressed as in (3.36), but now with $\Lambda \subset D$ and σ_n^a defined accordingly. The analogue of (3.38) would compare σ_n^a and a_m. Of course, $\sigma_n^a \le a_n$, $n \ge 1$. But no comparison $a_{cn} \le \sigma_n^a$, $n = 1, 2, \ldots$, is valid for any fixed constant $c \ge 1$. The reason is that adaptive algorithms do not create arbitrary functions in Σ_n^a. For example, the adaptive algorithm cannot have a partition with just one small dyadic interval; it automatically carries with it a certain entourage of intervals. We can explain this in more detail by using binary trees.

Consider any of the adaptive algorithms of the previous section. Given an $\epsilon > 0$, let $\mathcal{B}_\epsilon$ be the collection of all $I \in D$ such that $\mathcal{E}(I) > \epsilon$ (the collection of bad intervals). Then, whenever $I \in \mathcal{B}_\epsilon$, its parent is too. Thus $\mathcal{B}_\epsilon$ is a binary tree with root Ω. The set of dyadic intervals $\mathcal{G}_\epsilon$ is precisely the set of good intervals I (*i.e.*, $\mathcal{E}(I) \le \epsilon$) whose parent is bad. The inefficiency of the adaptive algorithm occurs when $\mathcal{B}_\epsilon$ contains a long chain of intervals

$I_1 \supset I_2 \supset \cdots \supset I_m$ with I_k the parent of I_{k+1} with the property that the other child of I_k is always good, $k = 1, \ldots, m-1$. This occurs, for example, when the target function f has a singularity at some point $x_0 \in I_m$ but is smooth otherwise. The partition $\mathcal{G}_\epsilon$ will contain one dyadic interval at each level (the sibling J_k of I_k). Using free knot partitions, we would zoom in faster on this singularity and thereby avoid this entourage of intervals J_k.

There are ways of modifying the adaptive algorithm to make it comparable to approximation from Σ_n^a, which we now briefly describe. If we are confronted with a long chain $I_0 \supset I_1 \supset \cdots \supset I_m$ of bad intervals from $\mathcal{B}_\epsilon$, the adaptive algorithm would place each of the sibling intervals J_k of I_k, $k = 0, \ldots, m$, into the good partition. We can decrease the number of intervals needed in the following way. We find the shortest subchain $I_0 = I_{j_0} \supset I_{j_1} \supset \cdots \supset I_{j_\ell} = I_m$ for which $\mathcal{E}(I_{j-1} \setminus I_j) < \epsilon$, $j = 1, \ldots, \ell$. Then, it is sufficient to use the intervals I_{j_i}, $i = 0, \ldots, \ell$, in place of the intervals J_k, $k = 0, \ldots, m$, in the construction of an approximant from Σ_n^a (see DeVore and Popov (1987) or Cohen, DeVore, Petrushev and Xu (1998) for a further elaboration on these ideas).

3.5. Wavelets: a first look; the Haar system

The two topics of approximating functions and representing them are closely related. For example, approximation by trigonometric sums is closely related to the theory of Fourier series. Is there an analogue in approximation by piecewise constants? The answer is yes. There are in fact several representations of a given function f using a basis of piecewise constant functions. The most important of these is the Haar basis, which we shall now describe.

Rather than simply introducing the Haar basis and giving its properties, we prefer to present this topic from the viewpoint of multiresolution analysis (MRA) since this is the launching point for the construction of wavelet bases, which we shall discuss in more detail in Section 7. Wavelets and multilevel methods are increasingly coming into favour in numerical analysis.

Let us return to the linear spaces $\mathcal{S}^1(\Delta_n)$ of piecewise constant functions on the partition of Ω with spacing $1/n$. We shall only need the case $n = 2^k$ and we denote this space by $\mathcal{S}_k := \mathcal{S}^1(\Delta_{2^k})$. The characteristic functions χ_I, $I \in D_k(\Omega)$, are a basis for $\mathcal{S}_k$. If we approximate well a smooth function f by a piecewise constant function $S = \sum_{I \in D_k} c_I \chi_I$ from $\mathcal{S}_k$, then the coefficients c_I will not change much: c_I will be close to c_J if I is close to J. We would like to take advantage of this fact to find a more compact representation for S. That is, we should be able to find a more favourable basis for $\mathcal{S}_k$ for which the coefficients of S are either zero or small.

The spaces $\mathcal{S}_k$ form a ladder: $\mathcal{S}_k \subset \mathcal{S}_{k+1}$, $k = 0, 1, \ldots$. We let $W_k := \mathcal{S}_{k+1} \ominus \mathcal{S}_k$ be the orthogonal complement of $\mathcal{S}_k$ in $\mathcal{S}_{k+1}$. This means that

W_k consists precisely of the functions in $w \in \mathcal{S}_{k+1}$ orthogonal to $\mathcal{S}_k$:

$$\int_\Omega w(x)S(x)\,\mathrm{d}x = 0, \quad \text{for all } S \in \mathcal{S}_k.$$

We then have

$$S_{k+1} = S_k \oplus W_k, \quad k = 0, 1, \ldots. \tag{3.39}$$

Thus W_k represents the *detail* that must be added to $\mathcal{S}_k$ in order to obtain $\mathcal{S}_{k+1}$.

The spaces W_k have a very simple structure. Consider, for example, $W := W_0$. Since $S_1 = S_0 + W_0$, and S_1 has dimension 2 and S_0 dimension 1, the space W_1 will be spanned by a single function from $\mathcal{S}_1$. Orthogonality gives us that this function is a nontrivial multiple of

$$H(x) := \chi_{[0,1/2)} - \chi_{[1/2,1)} = \begin{cases} 1, & 0 \le x < 1/2, \\ -1, & 1/2 \le x < 1. \end{cases} \tag{3.40}$$

H is called the Haar function. More generally, it is easy to see that W_k is spanned by the following (normalized) shifted dilates of H:

$$H_{j,k}(x) := 2^{k/2}H(2^k x - j), \quad j = 0, \ldots, 2^k - 1. \tag{3.41}$$

The function $H_{j,k}$ is a scaled version of H fitted to the interval $2^{-k}[j, j+1)$ which has $L_2(\Omega)$-norm one: $\|H_{j,k}\|_{L_2(\Omega)} = 1$.

From (3.39), we find

$$\mathcal{S}_m = \mathcal{S}_0 \oplus W_0 \oplus \cdots \oplus W_{m-1}. \tag{3.42}$$

It follows that χ_Ω together with the functions $H_{j,k}$, $j = 0, \ldots, 2^k - 1$, $k = 0, \ldots, m-1$, form an orthonormal basis for $\mathcal{S}_m$ which is, in many respects, better than the old basis χ_I, $I \in D_m$. But, before taking up that point, we want to see that we can take $m \to \infty$ in (3.42) and thereby obtain a basis for $L_2(\Omega)$.

It will be useful to have an alternative notation for the Haar functions $H_{j,k}$. Each j, k corresponds to the dyadic interval $I := 2^{-k}[j, j+1)$. We shall write

$$H_I := H_{j,k} = |I|^{-1/2}H(2^k \cdot -j). \tag{3.43}$$

From (3.42) we see that each $S \in \mathcal{S}_m$ has the representation

$$S = \langle S, \chi_\Omega\rangle\chi_\Omega + \sum_{I \in \cup_{0 \le k < m} D_k} \langle S, H_I\rangle H_I, \tag{3.44}$$

where

$$\langle f, g\rangle := \int_\Omega f(x)g(x)\,\mathrm{d}x \tag{3.45}$$

is the inner product in $L_2(\Omega)$.

Let P_m denote the orthogonal projector onto $\mathcal{S}_m$. Thus, $P_m f$ is the best $L_2(\Omega)$-approximation to f from $\mathcal{S}_m$. It is the unique element in $\mathcal{S}_m$ such that $f - P_m f$ is orthogonal to $\mathcal{S}_m$. Using the orthonormal basis of (3.44), we see that

$$P_m f = \langle f, \chi_\Omega \rangle \chi_\Omega + \sum_{I \in \cup_{0 \le k < m} D_k} \langle f, H_I \rangle H_I. \tag{3.46}$$

Since $\operatorname{dist}(f, \mathcal{S}_m)_{L_2(\Omega)} \to 0$, $m \to \infty$, we can take the limit in (3.46) to obtain

$$f = \langle f, \chi_\Omega \rangle \chi_\Omega + \sum_{I \in D} \langle f, H_I \rangle H_I \tag{3.47}$$

In other words, χ_Ω together with the functions H_I, $I \in D$, form an orthonormal basis, called the *Haar basis*, for $L_2(\Omega)$.

Some of the advantages of the Haar basis for $\mathcal{S}_m$ over the standard basis $(\chi_I,\ I \in D_m)$ are obvious. If we wish to increase our resolution of the target function by approximating from $\mathcal{S}_{m+1}$ rather than $\mathcal{S}_m$, we do not need to recompute our approximant. Rather, we merely add a layer of the decomposition (3.47) to the approximant corresponding to the wavelet space W_{m+1}. Of course, the orthogonality of W_m to S_m means that this new information is independent of our previous information about f. It is also clear that the coefficients of the basis function H_I, $I \in D_m$, tend to zero as $m \to \infty$. Indeed, we have

$$\|f\|_{L_2(\Omega)}^2 = |\langle f, \chi_\Omega \rangle|^2 + \sum_{I \in D} |\langle f, H_I \rangle|^2. \tag{3.48}$$

Therefore, this series converges absolutely.

3.6. n-term approximation: a second look

We shall next consider n-term approximation using the Haar basis. This is a special case of n-term wavelet approximation considered in more detail in Section 7.4. Let Σ_n^H denote the collection of all functions S of the form

$$S = c\chi_\Omega + \sum_{I \in \Lambda} c_I H_I, \tag{3.49}$$

where $\Lambda \subset D$ is a set of dyadic intervals with $\#\Lambda \le n$. As before, we let

$$\sigma_n^H(f)_p := \inf_{S \in \Sigma_n^H} \|f - S\|_{L_p(\Omega)} \tag{3.50}$$

be the error of n-term approximation.

We shall consider first the case of approximation in $L_2(\Omega)$ where the matter is completely transparent. In fact, in this case, in view of the norm equivalence (3.48), we see that a best approximation from Σ_n^H is given by

$$S = \langle f, \chi_\Omega \rangle \chi_\Omega + \sum_{I \in \Lambda} \langle f, H_I \rangle H_I, \tag{3.51}$$

where $\Lambda \subset D$ is a set corresponding to the n biggest Haar coefficients. Since there may be coefficients of equal absolute values, best approximation from $\sum_n^H$ is not necessarily unique.

Since we are dealing with an orthonormal system, we can apply the results of Section 2 to characterize the class of functions f which satisfy

$$\sigma_n^H(f)_2 \leq M n^{-\alpha}, \quad n = 1, 2, \ldots. \tag{3.52}$$

Namely, let $\gamma_n := \gamma_n(f)$ be the absolute of the nth largest Haar coefficient. It follows from the characterization (2.18) that, for any $\alpha > 0$, a function f satisfies (3.52) if and only if

$$\gamma_n(f) \leq \frac{M'}{n^{\alpha+1/2}}. \tag{3.53}$$

Moreover, the smallest constant M in (3.52) is equivalent (independently of f) to the smallest constant M' in (3.53).

It is interesting to note that the above characterization holds for any $\alpha > 0$; it is not necessary to assume that $\alpha \leq 1$. It is not apparent how the characterization (3.53) relates directly to the smoothness of f. We shall see later, when we develop n-term wavelet approximation in more detail, that, for $0 < \alpha < 1$, (3.53) is tantamount to requiring that f have α orders of smoothness in L_τ, where τ is defined by $1/\tau = \alpha + 1/2$. We recall our convention for interpreting smoothness spaces as points in the upper right quadrant of $\mathbb{R}^2$, as described in Section 3.2. The point $(1/\tau, \alpha)$ lies on the line with slope one which passes through $(1/2, 0)$ $(L_2(\Omega))$. Thus, the characterization of n-term Haar approximation (in $L_2(\Omega)$) is the same as the previous characterizations of free knot approximation.

The study of n-term Haar approximation in $L_2(\Omega)$ benefited greatly from the characterization of $L_2(\Omega)$ in terms of wavelet coefficients. The situation for approximation in $L_p(\Omega)$, $1 < p < \infty$, can also be treated, although the computation of $L_p(\Omega)$ norms is more subtle (see (7.27)). It turns out that a norm close to the L_p norm is given by

$$\|f\|_{B_p}^p := |\langle f, \chi_\Omega \rangle|^p + \sum_{I \in D} \|\langle f, H_I \rangle H_I\|_{L_p(\Omega)}^p, \tag{3.54}$$

which is known as the B_p norm. For approximation in the B_p norm, the theory is almost identical to $L_2(\Omega)$. Now, a best approximation from $\Sigma_n^H$is given by

$$S = \langle f, \chi_\Omega \rangle \chi_\Omega + \sum_{I \in \Lambda} \langle f, H_I \rangle H_I, \tag{3.55}$$

where $\Lambda \subset D$ is a set corresponding to the n biggest terms $\|\langle f, H_I\rangle H_I\|_{L_p(\Omega)}$. This selection procedure, to build the set Λ, depends on p because

$$\|H_I\|_{L_p(\Omega)} = |I|^{1/p-1/2}.$$

In other words, the coefficients are scaled depending on their dyadic level before we select the largest coefficients.

This same selection procedure works for approximation in L_p (DeVore, Jawerth and Popov 1992); however, now the proof is more involved and will be discussed in Section 7.4 when we treat the more general case of wavelets.

3.7. Optimal basis selection: wavelet packets

We have shown in Section 2 that, in the setting of a Hilbert space, it is a simple matter to determine a best n-term approximation to a target function f using elements of an orthonormal basis. A basis is good for f if the absolute value of the coefficients of f, when they are reordered according to decreasing size, tend rapidly to zero. We can increase our approximation efficiency by finding such a good basis for f. Thus, we may want to include in our approximation process a search over a given collection (usually called a *library*) of orthonormal bases in order to choose one which is good for our target function f. This leads to another degree of nonlinearity in our approximation process since now we have the choice of basis in addition to the choice of best n terms with respect to that basis. From a numerical perspective, however, we must be careful that this process can be implemented computationally. In other words, we cannot allow too many bases in our selection: our library of bases must be computationally implementable. In the case of piecewise constant approximation, such a library of bases was given by Coifman and Wickerhauser (1992) and is a special case of what are known as *wavelet packet libraries.*

We introduce some notation which will simplify our description of wavelet packet libraries. If g is a function from $L_2(\mathbb{R})$, we let

$$g_I(x) := |I|^{-1/2} g(2^n x - k), \quad I = 2^{-n}[k, k+1). \tag{3.56}$$

If g is supported on $\Omega = [0, 1)$, then g_I will be supported on the dyadic interval I. We also introduce the following scaling operators which appear in the construction of multiresolution analysis for the Haar function. For a function $g \in L_2(\mathbb{R})$, we define

$$A_0 g := g(2\cdot) + g(2\cdot -1)); \qquad A_1 g := g(2\cdot) - g(2\cdot -1)). \tag{3.57}$$

If g is supported on Ω, the functions $A_0 g$, $A_1 g$ are also supported on Ω and have the same L_2 norm as g. Also, the functions $A_0 g$ and $A_1 g$ are orthogonal, that is,

$$\int_\Omega A_0 g A_1 g = 0.$$

Let $\gamma_0 := \chi_\Omega$ and $\gamma_1 := H$ be the characteristic and Haar functions. They satisfy

$$\gamma_0 = A_0\gamma_0; \qquad \gamma_1 = A_1\gamma_0. \tag{3.58}$$

In the course of our development of wavelet packets we will apply the operators A_0 and A_1 to generate additional functions. It is most convenient to index these functions on binary strings b. Such a b is a string of 0s and 1s. For such a string b, let $b0$ be the new string obtained from b by appending 0 to the end of b and let $b1$ be the corresponding string obtained by appending 1 to the end of b. Then, we inductively define

$$\gamma_{b0} := A_0\gamma_b; \qquad \gamma_{b1} := A_1\gamma_b. \tag{3.59}$$

In particular, (3.58) gives that $\gamma_{00} := A_0\gamma_0 = \chi_\Omega$ and $\gamma_{01} := A_1\gamma_0 = H$. Note that there is redundancy in that two binary strings b and b' represent the same integer in base 2 if and only if $\gamma_b = \gamma_{b'}$.

We can now describe the wavelet packet bases for S_m with $m \geq 1$, a fixed integer. We associate to each binary string b its length $\#b$, and the space

$$\Gamma_b := \operatorname{span}\{(\gamma_b)_I : I \in D_{m-\#b}\}. \tag{3.60}$$

The functions $(\gamma_b)_I$ form an orthonormal basis for Γ_b. While the two functions γ_b and $\gamma_{b'}$ may be identical for $b \neq b'$, the subspaces Γ_b and $\Gamma_{b'}$ are not the same because b and b' will have different lengths. For any binary string b, we have

$$\Gamma_b = \Gamma_{b0} \oplus \Gamma_{b1}, \tag{3.61}$$

and the union of the two bases (given by (3.60)) for Γ_{b0} and Γ_{b1} give an alternative orthonormal basis for Γ_b.

The starting point of multiresolution analysis and our construction of the Haar wavelet was the decomposition $\mathcal{S}_m = \mathcal{S}_{m-1} \oplus W_{m-1}$ given in (3.42). In our new notation, this decomposition is

$$\Gamma_0 = \Gamma_{00} \oplus \Gamma_{01}. \tag{3.62}$$

In multiresolution analysis, the process is continued by decomposing $\mathcal{S}_{m-1} = \mathcal{S}_{m-2} \oplus W_{m-2}$ or, equivalently, $\Gamma_{00} = \Gamma_{000} \oplus \Gamma_{001}$. We take $W_{m-2} = \Gamma_{001}$ in our decomposition and continue. Our new viewpoint is that we can apply the recipe (3.57) to further decompose $\Gamma_{01} = W_{m-1}$ into two orthogonal subspaces as described in (3.61). Continuing in this way, we get other orthogonal decompositions of $\mathcal{S}_m$ and other orthonormal bases which span this space.

We can depict these orthogonal decompositions by a binary tree as given in Figure 4. Each node of the tree can be indexed by a binary string b. The number of digits k in b corresponds to its depth in the tree. Associated to b are the function γ_b and the space Γ_b, which has an orthonormal basis

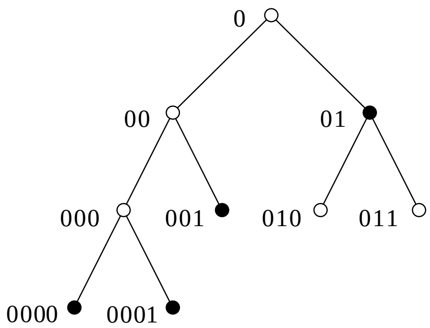

Fig. 4. The binary tree for wavelet packets

consisting of the functions $(\gamma_b)_I$, $I \in D_{m-k}$. If we move down and to the left from b we append digit 0 to b, while if we move down one level on the right branch we append digit 1. The tree stops when we reach level m.

The above construction generates many orthonormal bases of $\mathcal{S}_m$. We can associate each binary b to the dyadic interval I_b whose left end-point is $b/2\#b$ and whose length is $2^{-\#b}$, where $\#b$ is the number of digits in b (the level of b in the tree). If we take a collection B of such b, such that the I_b, $b \in B$, are a disjoint cover of Ω, then $\mathcal{S}_m = \Gamma_0 = \bigoplus_{b \in B} \Gamma_b$. The union of all the bases for these spaces form an orthonormal basis for $\mathcal{S}_m$. For example, in Figure 4, the solid nodes correspond to such a cover. The same story applies to any node b of the tree. The portion of the tree starting at b has the same structure as the entire tree and we obtain many bases for Γ_b by using interval decompositions of I_b as described above.

Several of the bases for $\mathcal{S}_m$ are noteworthy. By choosing just Ω in the interval decomposition of Ω, we obtain just the space Γ_0 and its basis $(\gamma_0)_I = \chi_I$, $I \in D_m$. The choice $B = \{00\cdots 0\} \cup \{01,\ 001, \ldots\}$ corresponds to the dyadic intervals $2^{-m}[0,1)$, $2^{-m+1}[1/2,1], \ldots,$ $[1/2,1)$ and gives the Haar basis. We can also take all the nodes at the lowest level (level m) of the tree. These nodes each correspond to spaces of dimension one. The basis obtained in this way is the Walsh basis from Fourier analysis.

It is important to note that we can efficiently compute the coefficients of a function $S \in \mathcal{S}_m$ with respect to all of the spaces Γ_b by using (3.57). For example, let γ_b be the generator of Γ_b. Then, $\Gamma_b = \Gamma_{b0} \oplus \Gamma_{b1}$. If $S \in \mathcal{S}_m$ and $c_{b,I} := \langle S, (\gamma_b)_I \rangle$, $I \in D_{m-k}$ are the coefficients of S with respect to these functions, then, for $I \in D_{m-k-1}$,

$$c_{b0,I} = \frac{1}{\sqrt{2}}(c_{b,I_0} + c_{b,I_1}), \qquad c_{b1,I} = \frac{1}{\sqrt{2}}(c_{b,I_0} - c_{b,I_1}), \tag{3.63}$$

where I_0 and I_1 are the left and right halves of I. Similarly, we can obtain the coefficients c_{b,I_0}, c_{b,I_1} from the coefficients $c_{b0,I}, c_{b1,I}$. Thus, for example, starting with the coefficients for the basis at the top (or bottom) of the tree, we can compute all other coefficients with $O(m2^m)$ operations.

For all numerical applications, the above construction is sufficient. One chooses m sufficiently large and considers all bases of $\mathcal{S}_m$ given as above. For theoretical reasons, however, one may want bases for $L_2(\Omega)$. This can be accomplished by letting $m \to \infty$ in the above depiction, thereby obtaining an infinite tree.

A typical adaptive basis selection algorithm, for approximating the target function f, chooses a coefficient norm that measures the spread of coefficients, and finds a basis that minimizes this norm. As we have seen in Section 2, n-term approximation efficiency using orthonormal bases is related to ℓ_τ norms of the coefficients. Thus, a typical algorithm would begin by fixing a sufficiently large value of m for the desired numerical accuracy, choosing $\tau > 0$, and finding a basis for the ℓ_τ norm, as we shall now describe.

If f is our target function, we let $S = P_m f$ be the orthogonal projection of f onto $\mathcal{S}_m$. The coefficients $\langle f, (\gamma_b)_I\rangle = \langle S, (\gamma_b)_I\rangle$ can be computed efficiently as described above. Let B be any orthonormal subcollection of the functions $(\gamma_b)_I$ and define

$$N_\tau(B) := N_\tau(f, B) := \sum_B |\langle f, (\gamma_b)_I\rangle|^\tau. \tag{3.64}$$

We want to find a basis B for Γ_0 which minimizes (3.64). To do this, we begin at the bottom of the tree and work our way up, at each step exchanging the current basis for a new one if the new basis gives a smaller N_τ.

For each node b at the bottom of the tree (*i.e.*, at level m), the space Γ_b has dimension one and has the basis $\{\gamma_b\}$. A node occurring at level $m-1$ corresponds to the space Γ_b. It has two bases from our collection. The first is $\{(\gamma_b)_I\}_{I\in D_1}$; the second is $\{\gamma_{b0}, \gamma_{b1}\}$. We compare these two bases and choose the one, which will be denoted by B_b, that minimizes $N_\tau(B)$. We do this for every node b at level $m-1$. We then proceed up the tree. If bases have been chosen for every node at level k, and if b is a node at level $k-1$, we compare $N_\tau(\{((\gamma_b)_I)_{I\in D_{k-1}}\})$ with $N_\tau(B_{b0} \cup B_{b1})$. The basis that minimizes N_τ is denoted by B_b and is our best basis for node b. At the conclusion, we shall have the best basis B_0 for node 0, that is, the basis which gives the smallest value of $N_\tau(B)$ among all wavelet packet bases for $\mathcal{S}_m$. This algorithm requires $O(m2^m)$ computations.

4. The elements of approximation theory

To move into the deeper aspects of nonlinear approximation, it will be necessary to call on some of the main tools of approximation theory. We have seen in the study of piecewise constant approximation that a prototypical theorem characterizes approximation efficiency in terms of the smoothness of the target function. For other methods of nonlinear approximation, it is not always easy to decide the appropriate measure of smoothness which characterizes approximation efficiency. There are, however, certain aids which make our search for this connection easier. The most important of these is the theory of interpolation of function spaces and the role of Jackson and Bernstein inequalities. This section will introduce the basics of interpolation theory and relate it to the study of approximation rates and smoothness. In the process, we shall engage three types of spaces: approximation spaces, interpolation spaces, and smoothness spaces. These three topics are intimately connected and it is these connections which give us insight on how to solve our approximation problems.

4.1. Approximation spaces

In our analysis of piecewise constant approximation, we have repeatedly asked the question: which functions are approximated at a given rate like $O(n^{-\alpha})$? It is time to put questions like this into a more formal framework. We shall consider the following general setting in this section. There will be a normed space $(X, \|\cdot\|_X)$, in which approximation takes place. Our approximants will come from spaces $X_n \subset X$, $n = 0, 1, \ldots$, and we introduce the approximation error

$$E_n(f)_X := \operatorname{dist}(f, X_n)_X := \inf_{g \in X_n} \|f - g\|_X. \tag{4.1}$$

In the case of linear approximation, n will usually be the dimension of X_n, or a quantity closely related to $\dim X_n$. In nonlinear approximation, n relates to the number of free parameters. For example, n might be the number of knots (breakpoints) in piecewise constant approximation with free knots. The X_n can be quite general spaces; in particular, they do not have to be linear. But we shall make the following assumptions (some only for convenience):

(i) $X_0 := \{0\}$
(ii) $X_n \subset X_{n+1}$
(iii) $aX_n = X_n$, $a \in \mathbb{R}$, $a \neq 0$
(iv) $X_n + X_n \subset X_{cn}$ for some integer constant $c \geq 1$ independent of n
(v) each $f \in X$ has a best approximation from X_n
(vi) $\lim_{n \to \infty} E_n(f)_X = 0$ for all $f \in X$.

Assumptions (iii), (iv), and (vi) are the most essential. The others can be eliminated or modified with a similar theory.

It follows from (ii) and (vi) that $E_n(f)_X$ monotonically decreases to 0 as n tends to ∞.

We wish to gather under one roof all functions which have a common approximation rate. In analogy with the results of the previous section, we introduce the space $\mathcal{A}^\alpha := \mathcal{A}^\alpha(X)$, which consists of all functions $f \in X$ for which

$$E_n(f)_X = O(n^{-\alpha}), \quad n \to \infty. \tag{4.2}$$

Our goal, as always, is to characterize $\mathcal{A}^\alpha$ in terms of something we know, such as a smoothness condition. It turns out that we shall sometimes need to consider finer statements about the decrease of the error $E_n(f)_X$. This will take the form of slight variants to (4.2), which we now describe.

Let $\mathbb{N}$ denote the set of natural numbers. For each $\alpha > 0$ and $0 < q < \infty$, we define the approximation space $\mathcal{A}_q^\alpha := \mathcal{A}_q^\alpha(X, (X_n))$ as the set of all $f \in X$ such that

$$|f|_{\mathcal{A}_q^\alpha} := \begin{cases} \left(\sum_{n=1}^\infty [n^\alpha E_n(f)_X]^q \frac{1}{n}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{n \geq 1} n^\alpha E_n(f)_X, & q = \infty, \end{cases} \tag{4.3}$$

is finite, and further define $\|f\|_{\mathcal{A}_q^\alpha} := |f|_{\mathcal{A}_q^\alpha} + \|f\|_X$. Thus, the case $q = \infty$ is the space $\mathcal{A}^\alpha$ described by (4.2). For $q < \infty$, the requirement for membership in $\mathcal{A}_q^\alpha$ gets stronger as q decreases:

$$\mathcal{A}_q^\alpha \subset \mathcal{A}_p^\alpha, \quad 0 < q < p \leq \infty.$$

However, all of these spaces correspond to a decrease in error like $O(n^{-\alpha})$.

Because of the monotonicity of the sequence $(E_n(f)_X)$, we have the equivalence

$$|f|_{\mathcal{A}_q^\alpha} \asymp \begin{cases} \left(\sum_{k=0}^\infty [2^{k\alpha} E_{2^k}(f)_X]^q\right)^{1/q}, & 0 < q < \infty, \\ \sup_{k \geq 0} 2^{k\alpha} E_{2^k}(f)_X, & q = \infty. \end{cases} \tag{4.4}$$

It is usually more convenient to work with (4.4) than (4.3).

The next sections will develop some general principles which can be used to characterize the approximation spaces $\mathcal{A}_q^\alpha$.

4.2. Interpolation spaces

Interpolation spaces arise in the study of the following problem of analysis. Given two spaces X and Y, for which spaces Z is it true that each linear operator T mapping X and Y boundedly into themselves automatically maps Z boundedly into itself? Such spaces Z are called *interpolation spaces* for the pair X, Y and the problem is to construct and, more ambitiously, to characterize the spaces Z. The classical result in this direction is the Riesz–Thorin theorem, which states that the spaces L_p, $1 < p < \infty$, are interpolation spaces for the pair L_1,L_∞ and the Calderón–Mitjagin theorem, which characterizes all the interpolation spaces for this pair as the rearrangement of invariant function spaces (see Bennett and Sharpley (1988)). There are two primary methods for constructing interpolation spaces Z: the complex method as developed by Calderón (1964a) and the real method of Lions and Peetre (see Peetre (1963)). We shall only need the latter in what follows.

Interpolation spaces arise in approximation theory in the following way. Consider our problem of characterizing the approximation spaces $\mathcal{A}^\alpha(X)$ for a given space X and approximating subspaces X_n. If we obtain information about $\mathcal{A}^\alpha(X)$ for a given value of α, is it possible to parlay that information into statements about other approximation spaces $\mathcal{A}^\beta(X)$, with $\beta \neq \alpha$? The answer is yes: we can interpolate this information. Using these ideas, we can usually characterize approximation spaces as interpolation spaces between X and a suitably chosen second space Y. Thus, our goal of characterizing approximation spaces gets reduced to that of characterizing certain interpolation spaces. Fortunately, much effort has been put into the problem of characterizing interpolation spaces, and characterizations (usually as smoothness spaces) are known for most classical pairs of spaces X, Y. Thus, our approximation problem is solved.

An example might motivate the reader. In our study of approximation by piecewise constants, we saw that $\mathrm{Lip}(1, L_p(\Omega))$ characterizes the functions which are approximated with order $O(n^{-1})$ in $L_p(\Omega)$ by linear approximation from $\mathcal{S}^1(\Delta_n)$. Interpolation gives that the spaces $\mathrm{Lip}(\alpha, L_p(\Omega))$ characterize the functions which are approximated with order $O(n^{-\alpha})$, $0 < \alpha < 1$. A similar situation exists in nonlinear approximation.

Our description of how to solve the approximation problem is a little unfair to approximation theory. It makes it sound as if we reduce the approximation problem to the interpolation problem and then call upon the interpolation theory for the final resolution. In fact, one can go both ways, that is, one can also think of characterizing interpolation spaces by approximation spaces. Indeed, this is often how interpolation spaces are characterized. Thus, both theories shed considerable light on the other, and this is the view we shall adopt in what follows.

As mentioned, we shall restrict our development to the real method of interpolation using the Peetre K-functional, which we now describe. Let X, Y be a pair of normed linear spaces. We shall assume that Y is continuously embedded in X ($Y \subset X$ and $\|\cdot\|_X \leq C\|\cdot\|_Y$). (There are a few applications in approximation theory where this is not the case and one can make simple modifications in what follows to handle those cases as well.) For any $t > 0$, we define the K-functional

$$K(f,t) := K(f,t;X,Y) := \inf_{g \in Y} \|f - g\|_X + t|g|_Y, \tag{4.5}$$

where $\|\cdot\|_X$ is the norm on X and $|\cdot|_Y$ is a semi-norm on Y. We shall also meet cases where $|\cdot|_Y$ is only a quasi-semi-norm, which means that the triangle inequality is replaced by $|g_1 + g_2|_Y \leq C(|g_1|_Y + |g_2|_Y)$ with an absolute constant C. To spare the reader, we shall ignore this distinction in what follows.

The function $K(f,\cdot)$ is defined on $\mathbb{R}_+$ and is monotone and concave (being the pointwise infimum of linear functions). Notice that, for each $t>0$, $K(f,t)$ describes a type of approximation. We approximate f by functions g from Y with the penalty term $t|g|_Y$. The role of the penalty term is paramount. As we vary $t>0$, we gain additional information about f.

K-functionals have many uses. As noted earlier, they were originally introduced as a means of generating interpolation spaces. To see that application, let T be a linear operator which maps X and Y into themselves with a norm not exceeding M in both cases. Then, for any $g \in Y$, we have $Tf = T(f-g) + Tg$ and therefore

$$K(Tf,t) \le \|T(f-g)\|_X + t|Tg|_Y \le M(\|f-g\|_X + t|g|_Y). \tag{4.6}$$

Taking an infimum over all g, we have

$$K(Tf,t) \le MK(f,t), \quad t>0. \tag{4.7}$$

Suppose further that $\|\cdot\|$ is a function norm defined for real-valued functions on $\mathbb{R}_+$. We can apply this norm to (4.7) and obtain

$$\|K(Tf,\cdot)\| \le M\|K(f,\cdot)\|. \tag{4.8}$$

Each function norm $\|\cdot\|$ can be used in (4.8) to define a space of functions (those functions for which the right side of (4.8) is finite) and this space will be an interpolation space. We shall restrict our attention to the most common of these, which are the θ,q norms. They are analogous to the norms we used in defining approximation spaces. If $0<\theta<1$ and $0<q\le\infty$, then the interpolation space $(X,Y)_{\theta,q}$ is defined as the set of all functions $f\in X$ such that

$$|f|_{(X,Y)_{\theta,q}} := \begin{cases} \left(\int_0^\infty [t^{-\theta}K(f,t)]^q \frac{\mathrm{d}t}{t}\right)^{1/q}, & 0<q<\infty, \\ \sup_{t>0} t^{-\theta}K(f,t), & q=\infty, \end{cases} \tag{4.9}$$

is finite.

The spaces $(X,Y)_{\theta,q}$ are interpolation spaces. The usefulness of these spaces depends on understanding their nature for a given pair (X,Y). This is usually accomplished by characterizing the K-functional for the pair. We shall give several examples of this in Sections 4.4–4.5.

Here is a useful remark which we shall have need for later. We can apply the θ,q method for generating interpolation spaces to any pair (X,Y). In particular, we can apply the method to a pair of θ,q spaces. The question is whether we get anything new and interesting. The answer is no: we simply get θ,q spaces of the original pair (X,Y). This is called the reiteration theorem of interpolation. Here is its precise formulation. Let $X' := (X,Y)_{\theta_1,q_1}$

and $Y' := (X, Y)_{\theta_2,q_2}$. Then, for all $0 < \theta < 1$ and $0 < q \le \infty$, we have

$$(X', Y')_{\theta,q} = (X, Y)_{\alpha,q}, \quad \alpha := (1-\theta)\theta_1 + \theta\theta_2. \tag{4.10}$$

We make two observations which can simplify the norm in (4.9). Firstly, using the fact that Y is continuously embedded in X, we obtain an equivalent norm by taking the integral in (4.9) over $[0, 1]$. Secondly, since $K(f, .)$ is monotone, the integral over $[0, 1]$ can be discretized. This gives that the norm of (4.9) is equivalent to

$$|f|_{(X,Y)_{\theta,q}} \asymp \begin{cases} \left(\sum_{k=0}^{\infty}[2^{k\theta}K(f, 2^{-k})]^q\right)^{1/q}, & 0 < q < \infty, \\ \sup_{k\ge 0} 2^{k\theta}K(f, 2^{-k}), & q = \infty \end{cases} \tag{4.11}$$

(see Chapter 6 of DeVore and Lorentz (1993) for details).

In this form, the definitions of interpolation spaces and approximation spaces are almost identical: we have replaced E_{2^k} by $K(f, 2^{-k})$. It should therefore come as no surprise that one space can often be characterized by the other. What is needed for this is a comparison between the error $E_n(f)$ and the K-functional K. Of course, this can only be achieved if we make the right choice of the space Y in the definition of K. But how can we decide what Y should be? This is the role of the Jackson and Bernstein inequalities given in the next subsection.

4.3. Jackson and Bernstein inequalities

In this subsection, we shall make a considerable simplification in the search for a characterization of approximation spaces and bring out fully the connection between approximation and interpolation spaces. We assume that X is the space in which approximation takes place and assume that we can find a positive number $r > 0$ and a second space Y continuously embedded in X for which the following two inequalities hold.

Jackson inequality: $E_n(f)_X \le Cn^{-r}|f|_Y, \quad f \in Y, \quad n = 1, 2, \ldots.$

Bernstein inequality: $|S|_Y \le Cn^r\|S\|_X, \quad S \in X_n, \quad n = 1, 2, \ldots.$

Whenever these two inequalities hold, we can draw a comparison between $E_n(f)_X$ and $K(f, n^{-r}, X, Y)$. For example, assume that the Jackson inequality is valid and let $g \in Y$ be such that

$$\|f - g\|_X + n^{-r}|g|_Y = K(f, n^{-r}).$$

(In fact we do not know of the existence of such a g, and so an ϵ should be added into this argument, but to spare the reader we shall not insist upon such precision in this survey.) If S is a best approximation to g from X_n, then

$$\begin{aligned} E_n(f)_X &\le \|f - S\|_X \le \|f - g\|_X + \|g - S\|_X \\ &\le \|f - g\|_X + Cn^{-r}|g|_Y \le CK(f, n^{-r}), \end{aligned} \tag{4.12}$$

where the last inequality makes use of the Jackson inequality.

By using the Bernstein inequality, we can reverse (4.12) in a certain weak sense (see Theorem 5.1 of Chapter 7 in DeVore and Lorentz (1993)). From this one derives the following relation between approximation spaces and interpolation spaces.

Theorem 1 If the Jackson and Bernstein inequalities are valid, then for each $0 < \gamma < r$ and $0 < q \leq \infty$ the following relation holds between approximation spaces and interpolation spaces

$$A_q^\gamma(X) = (X, Y)_{\gamma/r,q} \tag{4.13}$$

with equivalent norms.

Thus, Theorem 1 will solve our problem of characterizing the approximation spaces if we know two ingredients:

(i) an appropriate space Y for which the Jackson and Bernstein inequalities hold

(ii) a characterization of the interpolation spaces $(X, Y)_{\theta,q}$.

The first step is the difficult one from the viewpoint of approximation (especially in the case of nonlinear approximation). Fortunately, step (ii) is often provided by classical results in the theory of interpolation. We shall mention some of these in the next sections and also relate these to our examples of approximation by piecewise constants. But for now we want to make a very general and useful remark concerning the relation between approximation and interpolation spaces by stating the following elementary result of DeVore and Popov (1988*b*).

Theorem 2 For any space X and spaces X_n, as well as for any $r > 0$ and $0 \leq s \leq \infty$, the spaces X_n, $n = 1, 2, \ldots$, satisfy the Jackson and Bernstein inequalities for $Y = \mathcal{A}_s^r(X)$. Therefore, for any $0 < \alpha < r$ and $0 < q \leq \infty$, we have

$$\mathcal{A}_q^\alpha(X) = (X, \mathcal{A}_s^r(X))_{\alpha/r,q}. \tag{4.14}$$

In other words, the approximation family $\mathcal{A}_q^\alpha(X)$ is an interpolation family.

We also want to expand on our earlier remark that approximation can often be used to characterize interpolation spaces. We shall point out that, in certain cases, we can realize the K-functional by an approximation process.

We continue with the above setting. We say a sequence (T_n), $n = 1, 2, \ldots$, of (possibly nonlinear) operators, with T_n mapping X into X_n, provides *near best approximation* if there is an absolute constant $C > 0$ such that

$$\|f - T_n f\|_X \leq C E_n(f)_X, \quad n = 1, 2, \ldots.$$

We say this family is stable on Y if

$$|T_n f|_Y \le C|f|_Y, \quad n = 1, 2, \dots,$$

with an absolute constant $C > 0$.

Theorem 3 Let X, Y, (X_n) be as above and suppose that (X_n) satisfies the Jackson and Bernstein inequalities. Suppose further that the sequence of operators (T_n) provides near best approximation and is stable on Y. Then, T_n realizes the K-functional, that is,

$$\|f - T_n f\|_X + n^{-r}|T_n f|_Y \le CK(f, n^{-r}, X, Y)$$

with an absolute constant C.

For a proof and further results of this type, we refer the reader to Cohen, DeVore and Hochmuth (1997).

4.4. Interpolation for L_1, L_∞

The utility of the K-functional rests on our ability to characterize it and thereby characterize the interpolation spaces $(X, Y)_{\theta,q}$. Much effort was put forward in the 1970s and 1980s to establish such characterizations for classical pairs of spaces. The results were quite remarkable in that the characterizations that ensued were always in terms of classical entities that have a long-standing place in analysis. We shall give several examples of this. In the present section, we limit ourselves to the interpolation of Lebesgue spaces, which are classical to the theory. In later sections, we shall discuss interpolation of smoothness spaces, which are more relevant to our approximation needs.

Let us begin with the pair $L_1(A, \mathrm{d}\mu), L_\infty(A, \mathrm{d}\mu)$ with $(A, \mathrm{d}\mu)$ a given sigma-finite measure space. Hardy and Littlewood recognized the importance of the *decreasing rearrangement* f^* of a μ-measurable function f. The function f^* is a nonnegative, nonincreasing function defined on $\mathbb{R}_+$ which is equimeasurable with f:

$$\mu(f, t) := \mu\{x : |f(x)| > t\} = |\{s : f^*(s) > t\}|, \quad t > 0, \tag{4.15}$$

where we recall our notation for $|E|$ to denote the Lebesgue measure of a set E. The rearrangement f^* can be defined directly via

$$f^*(t) := \inf\{y : \mu(f, t) \le y\}. \tag{4.16}$$

Thus, f^* is essentially the inverse function to $\mu(f, t)$. We have the following beautiful formula for the K-functional for this pair (see Chapter 6 of DeVore and Lorentz (1993)):

$$K(f, t, L_1, L_\infty) = \int_0^t f^*(s)\,\mathrm{d}s, \tag{4.17}$$

which holds whenever $f \in L_1 + L_\infty$. From the fact that

$$\int_A |f|^p \,\mathrm{d}\mu = \int_0^\infty (f^*(s))^p \,\mathrm{d}s$$

it is easy to deduce from (4.17) the Riesz–Thorin theorem for this pair.

With the K-functional in hand, we can easily describe the (θ, q) interpolation spaces in terms of Lorentz spaces. For each $0 < p < \infty$, $0 < q \le \infty$, the Lorentz space $L_{p,q}(A, \mathrm{d}\mu)$ is defined as the set of all μ-measurable f such that

$$\|f\|_{L_{p,q}} := \begin{cases} (\int_0^\infty [t^{1/p} f^*(t)]^q \frac{\mathrm{d}t}{t})^{1/q}, & 0 < q < \infty, \\ \sup t^{1/p} f^*(t), & q = \infty, \end{cases} \tag{4.18}$$

is finite. Of course, the form of the integral in (4.18) is quite familiar to us. If we replace f^* by $\frac{1}{t}\int_0^t f^*(s)\,\mathrm{d}s = K(f,t)/t$ and use the Hardy inequalities (see Chapter 6 of DeVore and Lorentz (1993) for details) we obtain that

$$(L_1(A, \mathrm{d}\mu), L_\infty(A, \mathrm{d}\mu))_{1-1/p,q} = L_{p,q}(A, \mathrm{d}\mu), \quad 1 < p < \infty,\ 0 < q \le \infty. \tag{4.19}$$

Several remarks are in order. The space $L_{p,\infty}$ is better known as weak L_p and can be equivalently defined by the condition

$$\mu\{x : |f(x)| > y\} \le M^p y^{-p}. \tag{4.20}$$

The smallest M for which (4.20) is valid is equivalent to the norm in $L_{p,\infty}$.

The above results include the case when $\mathrm{d}\mu$ is purely atomic. This will be useful for us in what follows, in the following context. Let $\mathbb{N}$ be the set of natural numbers and let $\ell_p = \ell_p(\mathbb{N})$ be the collection of all sequences $x = (x(n))_{n\in\mathbb{N}}$ for which

$$\|x\|_{\ell_p} := \begin{cases} (\sum_{n=1}^\infty |x(n)|^p)^{1/p}, & 0 < p < \infty, \\ \sup_{n\in\mathbb{N}} |x(n)|, & p = \infty, \end{cases} \tag{4.21}$$

is finite. Then, $\ell_p(\mathbb{N}) = L_p(\mathbb{N}, \mathrm{d}\mu)$ with μ the counting measure. Hence, the above results apply. The Lorentz spaces in this case are denoted by $\ell_{p,q}$. The space $\ell_{p,\infty}$ (weak ℓ_p) consists of all sequences that satisfy

$$x^*(n) \le M n^{-1/p} \tag{4.22}$$

with $(x^*(n))$ the decreasing rearrangement of $(|x(n)|)$. This can equivalently be stated as

$$\#\{n : |x(n)| > y\} \le M^p y^{-p}. \tag{4.23}$$

The interpolation theory for L_p spaces applies to more than the pair (L_1, L_∞). We formulate this only for the spaces $\ell_{p,q}$ which we shall use later. For any $0 < p_1 < p_2 < \infty$, $0 < q_1, q_2 \le \infty$, we have

$$(\ell_{p_1,q_1}, \ell_{p_2,q_2})_{\theta,q} = \ell_{p,q}, \quad 1/p := \frac{1-\theta}{p_1} + \frac{\theta}{p_2}, \quad 0 < q \le \infty, \tag{4.24}$$

with equivalent norms. For $1 \le p_1, p_2 \le \infty$, this follows from (4.19) by using the reiteration theorem (4.10). The general case needs slight modification (see Bergh and Löfström (1976)).

Interpolation for the pair (L_1, L_∞) is rather unusual in that we have an exact identity for the K-functional. Usually we only get an equivalent characterization of K. One other case where an exact identity is known is interpolation between C and $\mathrm{Lip}\,1$, in which case

$$K(f, t; C, \mathrm{Lip}\,1) = \frac{1}{2}\bar{\omega}(f, 2t), \quad t > 0,$$

where ω is the modulus of continuity (to be defined in the next section) and $\bar{\omega}$ is its concave majorant (see Chapter 6 of DeVore and Lorentz (1993)).

4.5. Smoothness spaces

We have introduced various smoothness spaces in the course of discussing approximation by piecewise constants. In this section, we want to be a bit more systematic and describe the full range of smoothness spaces that we shall need in this survey. There are two important ways to describe smoothness spaces. One is through notions such as differentiability and moduli of smoothness. Most smoothness spaces were originally introduced into analysis in this fashion. A second way is to expand functions into a series of building blocks (for instance Fourier or wavelet) and describe smoothness as decay conditions on the coefficients in such expansions. That these descriptions are equivalent is at the heart of the subject. We shall give both descriptions. The first is given here in this section; the second in Section 7 when we discuss wavelet decompositions.

We begin with the most important and best known smoothness spaces, the Sobolev spaces. Suppose that $1 \le p \le \infty$ and $r > 0$ is an integer. If $\Omega \subset \mathbb{R}^d$ is a domain (for us this will mean an open, connected set), we define $W^r(L_p(\Omega))$ as the collection of all measurable functions f defined on Ω which have all their distributional derivatives $D^\nu f$, $|\nu| \le r$, in $L_p(\Omega)$. Here $|\nu| := |\nu_1| + \cdots + |\nu_d|$ when $\nu = (\nu_1, \ldots, \nu_d)$. The semi-norm for $W^r(L_p(\Omega))$ is defined by

$$|f|_{W^r(L_p(\Omega))} := \sum_{|\nu|=r} \|D^\nu f\|_{L_p(\Omega))}, \tag{4.25}$$

and their norm by $\|f\|_{W^r(L_p(\Omega))} := |f|_{W^r(L_p(\Omega))} + \|f\|_{L_p(\Omega)}$. Thus, Sobolev spaces measure smoothness of order r in L_p when r is a positive integer and $1 \le p \le \infty$. Their deficiency is that they do not immediately apply when r is nonintegral or when $p < 1$. We have seen several times already the need for smoothness spaces for these extended parameters when engaging nonlinear approximation.

We have seen in the Lipschitz spaces that one way to describe smoothness of fractional order is through differences. We have previously used only first differences; now we shall need differences of arbitrary order which we presently define. For $h \in \mathbb{R}^d$, let T_h denote the translation operator which is defined for a function f by $T_h f := f(\cdot + h)$ and let I denote the identity operator. Then, for any positive integer r, $\Delta_h^r := (T_h - I)^r$ is the rth *difference operator with step h*. Clearly $\Delta_h^r = \Delta_h(\Delta_h^{r-1})$. Also,

$$\Delta_h^r(f, x) := \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f(x + kh). \tag{4.26}$$

Here and later we use the convention that $\Delta_h^r(f, x)$ is defined to be zero when any of the points $x, \ldots, x + rh$ are not in Ω.

We can use Δ_h^r to measure smoothness. If $f \in L_p(\Omega)$, $0 < p \leq \infty$,

$$\omega_r(f, t)_p := \sup_{|h| \leq t} \|\Delta_h^r(f, \cdot)\|_{L_p(\Omega)} \tag{4.27}$$

is the rth *order modulus of smoothness of f in $L_p(\Omega)$*. In the case $p = \infty$, $L_\infty(\Omega)$ is replaced by $C(\Omega)$, the space of uniformly continuous functions on Ω. We always have that $\omega_r(f, t)_p \to 0$ monotonically as $t \to 0$. The faster this convergence to 0 the smoother is f.

We create smoothness spaces by bringing together all functions whose moduli of smoothness have a common behaviour. We shall particularly need this idea with the Besov spaces which are defined as follows. There will be three parameters in our description of Besov spaces. The two primary parameters are α, which gives the order of smoothness (for instance the number of derivatives), and p, which gives the L_p space in which smoothness is to be measured. A third parameter q, which is secondary to the two primary parameters, will allow us to make subtle distinctions in smoothness spaces with the same primary parameters.

Let $\alpha > 0$, $0 < p \leq \infty$, and $0 < q \leq \infty$. We take $r := [\alpha] + 1$ (the smallest integer larger than α). We say f is in the Besov space $B_q^\alpha(L_p(\Omega))$ if

$$|f|_{B_q^\alpha(L_p(\Omega))} := \begin{cases} \left(\int_0^\infty [t^{-\alpha}\omega_r(f, t)_p]^q \frac{\mathrm{d}t}{t}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{t>0} t^{-\alpha}\omega_r(f, t)_p, & q = \infty, \end{cases} \tag{4.28}$$

is finite. This expression defines the semi-norm on $B_q^\alpha(L_p(\Omega))$; the Besov norm is given by $\|f\|_{B_q^\alpha(L_p(\Omega))} := |f|_{B_q^\alpha(L_p(\Omega))} + \|f\|_{L_p(\Omega)}$. Here, we have complete analogy with the definitions (4.3) and (4.9) of approximation and interpolation spaces.

The Besov spaces give a full range of smoothness in that α can be any positive number, and p can range over $(0, \infty]$. As noted earlier, q is a secondary index which gives finer gradations of smoothness with the same primary indices.

We shall next make some further remarks which will help clarify Besov spaces, especially their relationship to other smoothness spaces such as the Sobolev and Lipschitz spaces. We assume from here on that the domain Ω is a Lipschitz domain (see Adams (1975)) – slightly weaker conditions on Ω suffice for most of the following statements.

We have taken r as the smallest integer larger than α. Actually, any choice of $r > \alpha$ will define the same space with an equivalent norm (see Chapter 2 of DeVore and Lorentz (1993)). If we take $\alpha < 1$ and $q = \infty$, the Besov space $B^\alpha_\infty(L_p(\Omega))$ is the same as $\mathrm{Lip}(\alpha, L_p(\Omega))$ with an identical semi-norm and norm. However, when $\alpha = 1$, we get a different space because the Besov space uses ω_2 in its definition but $\mathrm{Lip}(1, L_p(\Omega))$ uses ω_1. In this case, the Besov space is larger since $\omega_2(f,t)_p \le 2^{\max(1/p,1)}\omega_1(f,t)_p$. Sometimes $B^1_\infty(C(\Omega))$ is called the Zygmund space.

For the same reason that $\mathrm{Lip}\,1$ is not a Besov space, the Sobolev space $W^r_p(L_p(\Omega))$, $1 \le p \le \infty$, $p \ne 2$, r an integer, is not the same as the Besov space $B^r_\infty(L_p(\Omega))$. The Besov space is slightly larger. We could describe the Sobolev space $W^r_p(L_p(\Omega))$, $1 < p \le \infty$, by replacing ω_{r+1} by ω_r in the definition of $B^r_\infty(L_p(\Omega))$. When α is nonintegral, the fractional order Sobolev space $W^\alpha(L_p(\Omega))$ is defined to be $B^\alpha_p(L_p(\Omega))$. Two special cases are noteworthy. When $p = 2$, the Besov space $B^r_2(L_2(\Omega))$ is the same as the Sobolev space $W^r(L_2(\Omega))$; this is an anomaly that only holds for $p = 2$. The Lipschitz space $\mathrm{Lip}(1, L_1(\Omega))$ is the same as BV when Ω is an interval in $\mathbb{R}^1$. In higher dimensions, we use $\mathrm{Lip}(1, L_1(\Omega))$ as the definition of $\mathrm{BV}(\Omega)$; it coincides with some but not all of the many other definitions of BV.

Increasing the secondary index q in Besov spaces gives a larger space, *i.e.*,

$$B^\alpha_{q_1}(L_p(\Omega)) \subset B^\alpha_{q_2}(L_p(\Omega)), \quad q_1 < q_2.$$

However, the distinctions between these spaces are small.

The Sobolev embedding theorem gives much additional information about the relationship between Besov spaces with different values of the parameters. It is easiest to describe these results pictorially. As earlier, we identify a Besov space with primary indices p and α with the point $(1/p, \alpha)$ in the upper right quadrant of $\mathbb{R}^2$. The line with slope d passing through $(1/p, 0)$ is the demarcation line for embeddings of Besov spaces into $L_p(\Omega)$ (see Figure 5). Any Besov space with primary indices corresponding to a point above that line is embedded into $L_p(\Omega)$ (regardless of the secondary index q). Besov spaces corresponding to points on the demarcation line may or may not be embedded in $L_p(\Omega)$. For example the Besov spaces $B^\alpha_\tau(L_\tau(\Omega))$ with $1/\tau = \alpha/d + 1/p$ correspond to points on the demarcation line and they are embedded in $L_p(\Omega)$. Points below the demarcation line are never embedded in $L_p(\Omega)$.

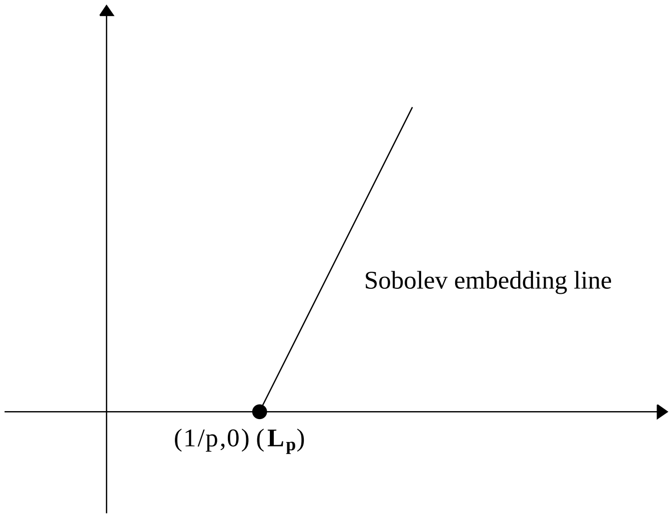

Fig. 5. Sobolev embedding

4.6. Interpolation of smoothness spaces

There is a relatively complete description of interpolation between Sobolev or Besov spaces. We shall point out the results most important for our later use.

Let us first consider interpolation between $L_p(\Omega)$ and a Sobolev space $W^r(L_p(\Omega))$. Interpolation for this pair appears often in linear approximation. One way to describe the interpolation spaces for this pair is to know its K-functional. The remarkable fact (proved in the case of domains by Johnen and Scherer (1977)) is that

$$K(f, t^r, L_p(\Omega), W^r(L_p(\Omega)) \asymp \omega_r(f,t)_p, \quad t > 0. \tag{4.29}$$

This brings home the point we made earlier that K-functionals can usually be described by some classical entity (in this case the modulus of smoothness). From (4.29), it is a triviality to deduce that

$$(L_p(\Omega), W^r(L_p(\Omega)))_{\theta,q} = B_q^{\theta r}(L_p(\Omega)), \quad 0 < \theta < 1,\ 0 < q \le \infty, \tag{4.30}$$

with equivalent norms. From the reiteration Theorem (4.10) for interpolation we deduce that, for $\alpha_1 < \alpha_2$ and any $0 < q_1, q_2 \le \infty$, we have for any $0 < \theta < 1$, $0 < q \le \infty$,

$$(B_{q_1}^{\alpha_1}(L_p(\Omega)), B_{q_2}^{\alpha_2}(L_p(\Omega)))_{\theta,q} = B_q^{\alpha}(L_p(\Omega)), \quad \alpha := (1-\theta)\alpha_1 + \theta\alpha_2. \tag{4.31}$$

We can also replace $B_{q_1}^{\alpha_1}(L_p(\Omega))$ by $L_p(\Omega)$ and obtain

$$(L_p(\Omega), B_r^{\alpha}(L_p(\Omega)))_{\theta,q} = B_q^{\theta\alpha}(L_p(\Omega)), \quad 0 < \theta < 1,\ 0 < q \le \infty, \tag{4.32}$$

for any $0 < r \le \infty$.

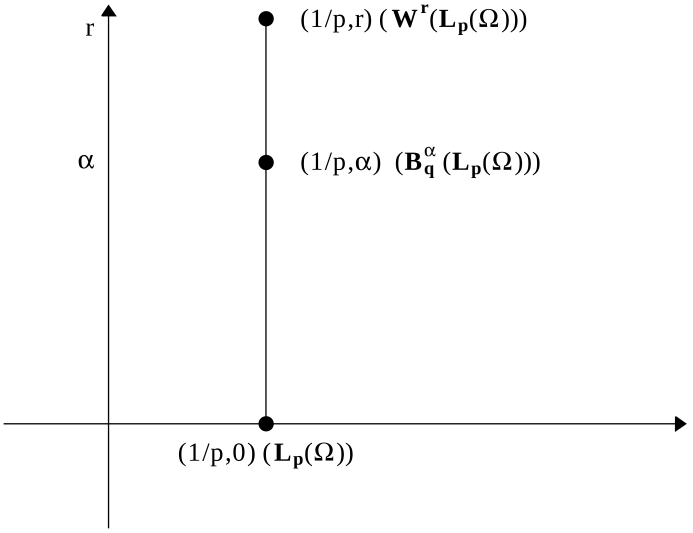

Fig. 6. Graphical interpretation of interpolation between $L_p(\Omega)$ and $W^r(L_p(\Omega))$

We can interpret these results pictorially as in Figure 6. The space $L_p(\Omega)$ corresponds to the point $(1/p, 0)$, and $W^r(L_p(\Omega))$ corresponds to the point $(1/p, r)$. Thus, (4.30) states that the interpolation spaces for this pair correspond to the Besov spaces on the (vertical) line segment connecting the points $(1/p, 0)$ and $(1/p, r)$. A similar picture interprets (4.31) and (4.32).

This pictorial interpretation is very instructive. When we want to interpolate between a pair of spaces (X, Y), we identify them with their corresponding points in the upper quadrant of $\mathbb{R}^2$. The points on the line segment connecting them are the interpolation spaces and, in fact, given the parameter θ, the interpolation space corresponds to the point on this line segment which divides the segment by the ratio $\theta : 1 - \theta$. However, care should be taken in this interpretation regarding the second parameter q, since it does not enter into the picture. In some cases, we can take any value of q, as is the case for the examples considered so far. However, in some cases that we shall see shortly, this interpretation only holds for certain q appropriately chosen.

Let us consider another example, which corresponds to interpolation in a case where the line segment is horizontal. DeVore and Scherer (1979) have shown that, if $1 \le p_1 < p_2 \le \infty$, then the (θ, p) interpolation between Sobolev spaces $W^r(L_{p_1}(\Omega))$ and $W^r(L_{p_2}(\Omega))$ gives Sobolev spaces $W^r(L_p(\Omega))$ when $\frac{1}{p} = \frac{1-\theta}{p_1} + \frac{\theta}{p_2}$, while changing θ, p into the more general θ, q gives the modified Sobolev spaces $W^r(L_{p,q}(\Omega))$ which use the Lorentz spaces $L_{p,q}(\Omega)$ in their definition (which we do not give).

There are characterizations for the θ, q interpolation spaces for many other pairs of Besov spaces (see Bergh and Löfström (1976) or Cohen, DeVore and Hochmuth (1997), for example). However, we shall restrict our further discussion to the following special case, which occurs in nonlinear approximation. We fix a value of $p \in (0, \infty)$ and consider the Besov spaces $B^\alpha_\tau(L_\tau(\Omega))$ where τ and α are related by

$$\frac{1}{\tau} = \frac{\alpha}{d} + \frac{1}{p}. \tag{4.33}$$

These spaces all correspond to points on the line segment with slope d passing through $(1/p, 0)$ (which corresponds to $L_p(\Omega)$). We have the following interpolation result for the pair $(L_p(\Omega), B^\alpha_\tau(L_\tau(\Omega)))$:

$$(L_p(\Omega), B^\alpha_\tau(L_\tau(\Omega)))_{\theta,q} = B^{\theta\alpha}_q(L_q(\Omega)), \quad \text{provided } \frac{1}{q} = \frac{\theta\alpha}{d} + \frac{1}{p}. \tag{4.34}$$

In other words, interpolating between two Besov spaces corresponding to points on this line, we get another Besov space corresponding to a point on this line provided we choose the secondary indices in a suitable way.

We shall obtain more information about Besov spaces and their interpolation properties in Section 7 when we discuss their characterization by wavelet decompositions.

5. Nonlinear approximation in a Hilbert space: a second look

Let us return to the example of approximation in a Hilbert space which began our discussion in Section 2. We continue with the discussion and notation of that section.

We have seen that for nonlinear (n-term) approximation in $\mathcal{H}$ we could characterize $\mathcal{A}^r_\infty((\mathcal{H}_n))$ for any $r > 0$ by the condition

$$\gamma_n(f) \le M n^{-r-1/2}, \tag{5.1}$$

with $\gamma_n(f)$ the rearranged coefficients. We now see that (5.1) states that the sequence $f_k := \langle f, \eta_k \rangle$ is in weak $\ell_{\tau(r)}$ ($\ell_{\tau(r),\infty}$), with $\tau(r)$ defined by

$$\tfrac{1}{\tau(r)} = r + \frac{1}{2},$$

and the smallest M for which (5.1) holds is equivalent to the weak ℓ_τ norm.

We can now characterize all of the approximation spaces $\mathcal{A}^\alpha_q(\mathcal{H})$ in terms of the coefficients f_k. Recall that Theorem 2 shows that, for any $r > 0$, the nonlinear spaces $\Sigma_n(\mathcal{H})$, satisfy Jackson and Bernstein inequalities for the space $Y := \mathcal{A}^r_\infty(\mathcal{H})$ and

$$\mathcal{A}^\alpha_q(\mathcal{H}) = (\mathcal{H}, \mathcal{A}^r_\infty(\mathcal{H}))_{\alpha/r,q}. \tag{5.2}$$

The mapping $f \to (f_k)$ is invertible and gives an isometry between $\mathcal{H}$ and $\ell_2(\mathbb{N})$ and also between $\mathcal{A}^r_\infty$ and $\ell_{\tau,\infty}(\mathbb{N})$. We can interpolate, and deduce that this mapping is an isometry between $\mathcal{A}^\alpha_q(\mathcal{H})$ and $\ell_{\tau(\alpha),q}(\mathbb{N})$ with τ defined by $1/\tau = \alpha + 1/2$. Hence, we have the following complete characterization of the approximation spaces for nonlinear n-term approximation.

Theorem 4 For nonlinear n-term approximation in a Hilbert space $\mathcal{H}$, a function f is in $\mathcal{A}^\alpha_q(\mathcal{H}))$ if and only if its coefficients are in $\ell_{\tau(\alpha),q}$, $\tau(\alpha) := (\alpha+1/2)^{-1}$, and $|f|_{\mathcal{A}^\alpha_q(\mathcal{H})} \asymp \|(f_k)\|_{\ell_{\tau(\alpha),q}}$.

6. Piecewise polynomial approximation

Now that we have the tools of approximation firmly in hand, we shall survey the main developments of nonlinear approximation, especially as they apply to numerical computation. We shall begin in this section with piecewise polynomial approximation. The reader should keep in mind the case of piecewise constant approximation that we used to motivate nonlinear approximation.

6.1. Local approximation by polynomials

As the name suggests, piecewise polynomial approximation pieces together local polynomial approximants. Therefore, we need to have a good understanding of local error estimates for polynomial approximation. This is an old and well-established chapter in approximation and numerical computation, which we shall briefly describe in this section.

For each positive integer r, we let $\mathcal{P}_r$ denote the space of polynomials in d variables of total degree $< r$ (polynomials of order r). Let $0 < p \le \infty$ and let I be a cube in $\mathbb{R}^d$. If $f \in L_p(I)$, the local approximation error is defined by

$$E_r(f, I)_p := \inf_{P \in \mathcal{P}_r} \|f - P\|_{L_p(I)}. \tag{6.1}$$

The starting point for estimating the efficiency of piecewise polynomial approximation in L_p is to have good estimates for $E_r(f, I)_p$. Perhaps the simplest of these is the estimate

$$E_r(f, I)_p \le C|I|^{r/d} |f|_{W^r(L_p(I))}, \tag{6.2}$$

which holds for $1 \le p \le \infty$, $|\cdot|_{W^r(L_p(I))}$ the Sobolev semi-norm of Section 4.5, and the constant C depending only on r. This is sometimes known as the Deny–Lions lemma in numerical analysis. There are several proofs of this result available in the literature (see, for instance, Adams (1975)), usually by constructing a bounded projector from L_p onto $\mathcal{P}_r$. It can also be proved indirectly (see DeVore and Sharpley (1984)).

The estimate (6.2) remains valid when I is replaced by a more general domain. Suppose, for example, that $\mathcal{O}$ is a domain that satisfies the uniform cone condition (see Adams (1975)) and is contained in a cube I with $|I|^{1/d} \le C\mathrm{diam}(\mathcal{O})$. If $f \in W^r(L_p(\mathcal{O}))$, then it can be extended to a function on I with comparable norm (Adams (1975) or DeVore and Sharpley (1984)). Applying (6.2) for I we deduce its validity on $\mathcal{O}$ with a constant C now depending on r and $\mathcal{O}$. We shall use this in what follows for polyhedral domains. The constant C then depends on r and the smallest angle in $\mathcal{O}$. Similar remarks apply to the other estimates for $E_r(f, I)_p$ that follow.

Using the ideas of interpolation introduced in Section 4 (see (4.29)), one easily derives from (6.2) that

$$E_r(f, I)_p \le C_r \omega_r(f, |I|, I)_p, \tag{6.3}$$

with ω_r the rth order modulus of smoothness of f introduced in Section 4.5. This is called Whitney's theorem in approximation and this estimate is equally valid in the case $p < 1$. The advantage of (6.3) over (6.2) is that it applies to any $f \in L_p(I)$ and it also implies (6.2) because of elementary properties of ω_r.

Whitney's estimate is not completely satisfactory when it is necessary to add local estimates over varying cubes I. A more suitable form is obtained by replacing $\omega_r(f, |I|, I)_p$ by

$$w_r(f, I)_p := \left(\frac{1}{|I|} \int_{|s| \le |I|^{1/d}} \int_I |\Delta_s^r(f, x)|^p \,\mathrm{d}x \,\mathrm{d}s \right)^{1/p}. \tag{6.4}$$

Then, we have (see, for instance, DeVore and Popov (1988*a*))

$$E_r(f, I)_p \le C_r w_r(f, I)_p, \tag{6.5}$$

which holds for all $r \ge 1$ and all $0 < p \le \infty$ (with the obvious change in norms for $p = \infty$).

It is also possible to bound $E_r(f, I)_p$ in terms of smoothness measured in spaces L_q, $q \ne p$. Such estimates are essentially embedding theorems and are important in nonlinear approximation. For example, in analogy with (6.2), we have for each $1 \le q, p \le \infty$ and $r > d(1/q - 1/p)_+$,

$$E_r(f, I)_p \le C_r |I|^{r/d + 1/p - 1/q} |f|_{W^r(L_q(I))}. \tag{6.6}$$

We shall sketch a simple idea for proving such estimates, which is at the heart of proving embedding theorems. We consider $q \le p$ (the other case is trivial). It is enough to prove (6.6) in the case $I = [0, 1]^d$ since it follows for other cubes by a linear change of variables. For each dyadic cube $J \subset I$, let P_J be a polynomial in $\mathcal{P}_r$ that satisfies

$$\|f - P_J\|_{L_q(J)} \le E_r(f, J)_q,$$

and define $S_k := \sum_{J\in D_k(I)} P_J\chi_J$, where $D_k(I)$ is the collection of all dyadic subcubes of I of side length 2^{-k}. Then, $S_0 = P_I$ and $S_k \to f$ in $L_p(I)$. Therefore,

$$E_r(f,I)_p \le \|f-P_I\|_{L_p(I)} \le \sum_{k=0}^{\infty} \|S_{k+1}-S_k\|_{L_p(I)}. \tag{6.7}$$

Now, for each polynomial $P \in \mathcal{P}_r$ and each cube J, we have $\|P\|_{L_p(J)} \le C|J|^{1/p-1/q}\|P\|_{L_q(J)}$ with the constant depending only on r (see Lemma 3.1 of DeVore and Sharpley (1984) for the simple proof). From this, it follows that

$$\begin{aligned}\|S_{k+1}-S_k\|^p_{L_p(I)} &= \sum_{J\in D_{k+1}(I)} \|S_{k+1}-S_k\|^p_{L_p(J)} \\ &\le C2^{-kd(1-p/q)} \sum_{J\in D_{k+1}(I)} \|S_{k+1}-S_k\|^p_{L_q(J)}.\end{aligned}$$

Now on J, we have $S_{k+1}-S_k = P_{J'}-P_J$ where J' is the parent of J. We write $P_{J'}-P_J = P_{J'}-f+f-P_J$ and use (6.2) (with p replaced by q) on each difference to obtain

$$\begin{aligned}\|S_{k+1}-S_k\|^p_{L_p(I)} &\le C2^{-kd(rp/d+1-p/q)} \sum_{J\in D_{k+1}(I)} |f|^p_{W^r(L_q(J'))} \\ &\le C2^{-kd(rp/d+1-p/q)} \left(\sum_{J\in D_{k+1}(I)} |f|^q_{W^r(L_q(J'))}\right)^{p/q} \\ &= C2^{-kd(rp/d+1-p/q)} |f|^p_{W^r(L_q(I))}.\end{aligned}$$

Here we used the facts that $\|\cdot\|_{\ell_p} \le \|\cdot\|_{\ell_q}$ if $q \le p$ and that a point $x \in I$ appears at most 2^d times in a cube J', as J runs over the cubes in $D_k(I)$. If we use this estimate in (6.7), we arrive at (6.6).

We can also allow $q < 1$ and nonintegral r in (6.6) if we use the Besov spaces. Namely, if $r > 0$ satisfies $r \ge d(1/q-1/p)_+$, then

$$E_r(f,I)_p \le C_r|I|^{r/d+1/p-1/q}|f|_{B^r_q((L_q(I))}. \tag{6.8}$$

Notice that we can allow $r/d+1/p-1/q = 0$ in (6.8), which corresponds to the embedding of $B^r_q(L_q(I))$ into $L_p(I)$. The case $r/d > (1/q-1/p)_+$ can be proved as above using the set subadditivity of $|\cdot|^q_{B^r_q(L_q(J))}$. For proofs of these results for Besov spaces see DeVore and Popov (1988a).

Finally, as we have remarked earlier, by using extensions, these results can be established for more general domains such as domains with a uniform cone condition. In particular, for any polyhedron $\mathcal{C}$, we have

$$E_r(f, \mathcal{C})_p \le C_r \mathrm{diam}(\mathcal{C})^{r/d+1/p-1/q} |f|_{W^r(L_q(\mathcal{C}))}, \tag{6.9}$$

with the constant depending only on r, d, the number of vertices of $\mathcal{C}$, and the smallest angle in $\mathcal{C}$. Similarly, we have the extension of (6.8) to polyhedra.

6.2. Piecewise polynomial approximation: the linear case

For the purpose of orienting the results on nonlinear approximation which follow, we shall in this section consider approximation by piecewise polynomials on fixed partitions. These results will be the analogue of approximation by piecewise constants on uniform partitions given in Section 3.1. For convenience, we shall consider approximation on the unit cube $\Omega := [0, 1]^d$. The following results can be established for more general domains by using extension theorems similar to what we have mentioned earlier in this section.

By a *partition* of Ω, we mean a finite collection $\Delta := \{\mathcal{C}\}$ of polyhedrons $\mathcal{C}$ which are pairwise disjoint and union to Ω. Given such a collection, we define the *partition diameter*

$$\mathrm{diam}(\Delta) := \max_{\mathcal{C}} \mathrm{diam}(\mathcal{C}). \tag{6.10}$$

We assume that the number of vertices of each cell $\mathcal{C}$ is bounded independently of $\mathcal{C} \in \Delta$.

Let $\mathcal{S}^r(\Delta)$ denote the space of piecewise polynomials of order r relative to Δ. That is, a function S is in $\mathcal{S}^r(\Delta)$ if and only if it is a polynomial of order r on each cell $\mathcal{C} \in \Delta$. For $0 < p \le \infty$, we let

$$s_\Delta(f)_p := \inf_{S \in \mathcal{S}^r(\Delta)} \|f - S\|_{L_p(\Omega)}. \tag{6.11}$$

We shall fix $1 \le p \le \infty$ and estimate $s_\Delta(f)_p$. A similar analysis holds for $p < 1$ with Sobolev norms replaced by Besov norms.

We assume that each cell $\mathcal{C}$ is contained in a cube $J \subset I$ with $|J|^{1/d} \le C\mathrm{diam}(\mathcal{C})$ with C depending only on c_Δ. Hence, by extending f to this cube (if it is not already defined there) we see that, for each $\mathcal{C} \in \Delta$, there is a polynomial $P_{\mathcal{C}} \in \mathcal{P}_r$ which satisfies (6.9):

$$\|f - P_{\mathcal{C}}\|_{L_p(\mathcal{C})} \le C \mathrm{diam}(\Delta)^r |f|_{W^r(L_p(\mathcal{C}))}. \tag{6.12}$$

If we raise the estimates in (6.12) to the power p (in the case $p < \infty$) and add them, we arrive at

$$s_\Delta(f)_p \le C \mathrm{diam}\, \Delta^r |f|_{W^r(L_p(\Omega))}. \tag{6.13}$$

Of course, (6.12) is well known in both approximation and numerical circles. It is the proper form for numerical estimates based on piecewise polynomials of order r. It is the *Jackson inequality* for this type of approximation. By interpolation (as described in Section 4.2), we obtain the following estimate

$$s_\Delta(f)_p \leq C\omega_r(f, \operatorname{diam}\Delta)_p, \tag{6.14}$$

where $\omega_r(f, \dot{)}_p = \omega_r(f, \cdot, \Omega)_p$ is the rth order modulus of smoothness of f in $L_p(\Omega)$ as introduced in Section 4.5. The advantage of (6.14) is that it does not require that f is in $W^r(L_p(\Omega))$ and in fact applies to any $f \in L_p(\Omega)$. For example, if $f \in \operatorname{Lip}(\alpha, L_p(\Omega))$, then (6.14) implies

$$s_\Delta(f)_p \leq C|f|_{\operatorname{Lip}(\alpha, L_p(\Omega))}|\operatorname{diam}\Delta|^\alpha. \tag{6.15}$$

We would now like to understand to what extent estimates like (6.15) are best possible. It is not difficult to prove that, if $f \in L_p(\Omega)$ is a function for which

$$s_\Delta(f)_p \leq M|\operatorname{diam}\Delta|^\alpha \tag{6.16}$$

holds *for every partition* Δ, then $f \in \operatorname{Lip}(\alpha, L_p(\Omega))$ and the smallest M for which (6.16) holds is equivalent to $|f|_{\operatorname{Lip}(\alpha, L_p(\Omega))}$. Indeed, for each $h \in \mathbb{R}^d$ and each $x \in \Omega$ such that the line segment $[x, x + rh) \subset \Omega$, there is a partition Δ with $\operatorname{diam}(\Delta) \leq |h|$ and $\operatorname{dist}(x, \partial\mathcal{C}) \geq \operatorname{const}|h|$ for every $\mathcal{C} \in \Delta$. This allows an estimate for $|\Delta_h^r(f, x)|$ by using ideas similar to the inverse estimates for piecewise constant approximation given in Section 3.1.

We note that the direct and inverse theorems relating approximation order to smoothness take the same form as those in Section 3.1. Using our interpretation of smoothness spaces given in Figure 6, we see that the approximation spaces for this form of linear approximation correspond to points on the vertical line segment joining $(1/p, 0)$ (L_p) to $(1/p, r)$ $(\operatorname{Lip}(r, L_p)$. Thus the only distinction from the piecewise constant case considered in Section 3.1 is that we can allow α to range over the larger interval $[0, r]$ because we are using piecewise polynomials of order r. Also, note that to achieve approximation order $O(n^{-\alpha})$ we would need spaces $\mathcal{S}^r(\Delta_n)$ of linear space dimension $\approx n^d$, that is, we have the curse of dimensionality.

More generally, if we only know (6.16) for a specific sequence of partitions (Δ_n), we can still prove that $f \in \operatorname{Lip}(\alpha, L_p(\Omega))$ provided the partitions mix sufficiently so that each x falls in the middle of sufficiently many $\mathcal{C}$. We do not formulate this precisely but refer readers to Section 2 of Chapter 12 of DeVore and Lorentz (1993) for a precise formulation in the univariate case.

Mixing conditions are not valid in most numerical settings. Indeed, the typical numerical case is where approximation takes place from a sequence $\mathcal{S}^r(\Delta_n)$, where Δ_n is a refinement of Δ_{n-1}. This means that the spaces are nested: $\mathcal{S}^r(\Delta_{n-1}) \subset \mathcal{S}^r(\Delta_n)$, $n = 1, 2, \ldots$. In this case, one can prove the inverse theorems only for a smaller range of α. It is easy to see that restrictions are needed on α. For example, functions f in $\mathcal{S}^r(\Delta_n)$ will be approximated exactly for $m \geq n$. But functions in $\mathcal{S}^r(\Delta_n)$ do not have much smoothness because they are discontinuous across the faces of the partition. This can be remedied by considering approximation by elements of $\mathcal{S}^r(\Delta_n)$ which have additional smoothness across the faces of the partition. We do not formulate inverse theorems in this case but refer the reader to Section 3 of Chapter 12 in DeVore and Lorentz (1993) where similar univariate results are proved.

We should mention, however, that considering splines with smoothness brings out new questions concerning direct estimates of approximation like (6.12). It is not easy to understand the dimension of spaces of smooth multivariate piecewise polynomials, let alone their approximation power (see Jia (1983)).

As the reader can now see, there are still interesting open questions concerning the approximation power of splines on general partitions, which relate the smoothness of the splines to the approximation power. These are difficult problems and have to a large extent been abandoned with the advent of box splines and, later, wavelets. These two developments shifted the viewpoint of spline approximation away from partitions and more toward the spanning functions. We shall get into this topic more in Section 7 when we discuss wavelet approximation.

6.3. Free knot piecewise polynomial approximation

To begin our development of nonlinear approximation by piecewise polynomials we shall consider the case of approximating a univariate function f defined on $\Omega = [0, 1]$ by piecewise polynomials of fixed order r. The theory here is the analogue of piecewise constants discussed in Section 3.2.

Let the natural number r be fixed and for each $n = 1, 2, \ldots$, let $\Sigma_n := \Sigma_{n,r}$ be the space of piecewise polynomials of order r with n pieces on Ω. Thus, for each element $S \in \Sigma_n$ there is a partition Λ of Ω consisting of n disjoint intervals $I \subset \Omega$ and polynomials $P_I \in \mathcal{P}_r$ such that

$$S = \sum_{I \in \Lambda} P_I \chi_I. \tag{6.17}$$

For each $0 < p \leq \infty$, we define the error of approximation

$$\sigma_n(f)_p := \sigma_{n,r}(f)_p := \inf_{S \in \Sigma_{n,r}} \|f - S\|_{L_p(\Omega)}. \tag{6.18}$$

The case $p = \infty$ is sufficiently different that we shall restrict our discussion to the case $p < \infty$ and refer the reader to DeVore and Lorentz (1993) or Petrushev and Popov (1987) for the case $p = \infty$.

We can characterize the functions f that can be approximated with an order like $O(n^{-\alpha})$. We recall the approximation spaces

$$\mathcal{A}_q^\alpha(L_p(\Omega)) = \mathcal{A}_q^\alpha(L_p(\Omega), (\Sigma_n)).$$

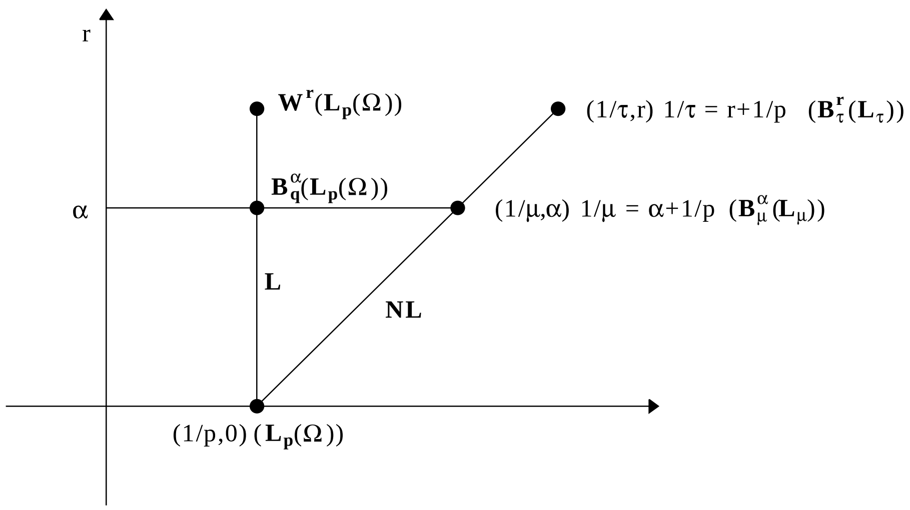

Fig. 7. Nonlinear approximation in L_p

According to the theory in Section 4, we can characterize these approximation spaces if we establish Jackson and Bernstein inequalities for this type of approximation. We fix the space $L_p(\Omega)$ in which approximation is going to take place. The space Y will be the Besov space $B^r_\tau(L_\tau(\Omega))$, $1/\tau = r + 1/p$ which was defined in Section 4.5. To understand this space, we return to our picture of smoothness spaces in Figure 7. The space $L_p(\Omega)$ of course corresponds to the point $(1/p, 0)$. The space $B^r_\tau(L_\tau(\Omega))$ corresponds to the point $(1/\tau, r)$, which lies on the line with slope one that passes through $(1/p, 0)$. As we have noted several times before, this line corresponds to the limiting case of the Sobolev embedding theorem. Thus, we are in complete analogy with the case of piecewise constant approximation described in Section 3.2.

The following inequalities were established by Petrushev (1988)

$$\sigma_n(f)_p \le Cn^{-r}|f|_{B^r_\tau(L_\tau(\Omega))} \tag{6.19}$$

$$|S|_{B^r_\tau(L_\tau(\Omega))} \le Cn^r\|f\|_{L_p(\Omega)} \tag{6.20}$$

with the constants C depending only on r. The first of these is the Jackson inequality and the second the companion Bernstein inequality.

Let us say a few words about how one proves these inequalities, since the techniques for doing so appear often in nonlinear approximation. To prove the Jackson inequality, for each $f \in B^\alpha_\tau(L_\tau(\Omega))$, we must find a favourable partition of Ω into n disjoint intervals. This is done by balancing $\Phi(I) := |f|^\tau_{B^\alpha_\tau(L_\tau(I))}$. The key here is that, with a proper renormalization of the Besov norm, Φ is set subadditive. Thus, we can find intervals I_j, $j = 1, \dots, n$, so that $\Phi(I_j) = \Phi(\Omega)/n$. This gives our desired partition. We then use (6.8) to bound the local approximation error on each I_j, add these and arrive at (6.19) (see Chapter 12 of DeVore and Lorentz (1993) for more details). Therefore, as was the case in our introduction of nonlinear approximation by piecewise constants, we find our optimal partitions by a balancing suitable set function, in the present case Φ.

The proof of the Bernstein inequality is also very instructive. If $S \in \Sigma_n$, then $S = \sum_{I\in\Lambda} \gamma_I$ where Λ is a partition of Ω into n intervals and $\gamma_I = P_I\chi_I$ with $P_I \in \mathcal{P}_r$. For each such γ_I, it is not difficult to calculate its Besov norm and find

$$|\gamma_I|_{B^\alpha_\tau(L_\tau(\Omega))} \le C\|\gamma_I\|_{L_p(I)}, \tag{6.21}$$

with C an absolute constant. Then, using the subadditivity of $|\cdot|^\tau_{B^\alpha_\tau(L_\tau(\Omega))}$, we find that

$$\begin{aligned} |S|^\tau_{B^\alpha_\tau(L_\tau(\Omega))} &\le \sum_{I\in\Lambda} |\gamma_I|^\tau_{B^\alpha_\tau(L_\tau(\Omega))} \\ &\le \sum_{I\in\Lambda} \|\gamma_I\|^\tau_{L_p(\Omega)} \\ &\le Cn^{1-\tau/p}\left(\sum_{I\in\Lambda}\|\gamma_I\|^p_{L_p(\Omega)}\right)^{\tau/p} = Cn^{\alpha\tau}\|S\|^\tau_{L_p(\Omega)}. \end{aligned}$$

With the Jackson and Bernstein inequalities in hand, we can now refer to our general theory in Section 4 and obtain the following characterization of the approximation spaces: for each $0<\alpha<r$, $0<q\le\infty$, $0<p<\infty$,

$$\mathcal{A}^\alpha_q(L_p(\Omega)) = (L_p(\Omega), B^r_\tau(L_\tau(\Omega)))_{\alpha/r,q}. \tag{6.22}$$

Therefore, we have a solution to our problem of characterizing the approximation spaces $A^\alpha_q(L_p(\Omega))$ to the extent that we understand the interpolation spaces appearing in (6.22). Fortunately, we know a lot about these interpolation spaces. For example, for each $0<\alpha<r$, there is one value of q for which this interpolation space is a Besov space. Namely, if $1/q = \alpha + 1/p$, then

$$\mathcal{A}^\alpha_q(L_p(\Omega)) = (L_p(\Omega), B^r_\tau(L_\tau(\Omega))_{\alpha/r,q} = B^\alpha_q(L_q(\Omega)). \tag{6.23}$$

For other values of q these interpolation spaces can be described in various ways. We defer a discussion of this until we treat nonlinear wavelet approximation where these interpolation spaces will reappear.

Returning to our picture of smoothness spaces, we see that the approximation spaces $\mathcal{A}^\alpha_q(L_p(\Omega))$ correspond to the point $(1/\tau,\alpha)$ with $1/\tau = \alpha + 1/p$. Thus, these spaces lie on the line with slope one passing through $(1/p, 0)$. In other words, we have the same interpretation as in nonlinear approximation by piecewise constants except that now α can range over the larger interval $(0, r)$ corresponding to the order r of the piecewise polynomials.

We have emphasized that the Besov spaces $B^\alpha_\tau(L_\tau(\Omega))$, $1/\tau = \alpha + 1/p$, which occur in characterizing free knot spline approximation, lie on the demarcation line in the Sobolev embedding theorem. This is an indication that these spaces are quite large when compared to the Besov spaces $B^\alpha_q(L_p(\Omega))$ which appear in characterizing linear approximation. Some examples might further drive this point home. Any function f which is a piecewise polynomial (with a finite number of pieces) is in all of these spaces, that is, we can take α arbitrarily large. Indeed, f can be approximated exactly once n and r are large enough and hence this result follows from (6.22). A simple argument shows that this remains true for any piecewise analytic function f. Hence, any such function can be approximated to accuracy $O(n^{-\alpha})$ for any $\alpha > 0$ with nonlinear piecewise polynomial approximation. Another instructive example is the function $f(x) = x^\gamma$, $\gamma > -1/p$ (so that $f \in L_p(\Omega)$). This function satisfies (see de Boor (1973))

$$\sigma_{n,r}(f)_p = O(n^{-r}).$$

This can be proved by balancing the approximation errors.

6.4. Free knots and free degree

There are many variants of piecewise polynomial approximation. One of the most important is to allow not only the partition to vary with f but also the orders (degrees) of the polynomial pieces. Approximation of this type occurs in the *h*-*p* method in FEM which has been introduced and studied by Babuška and his collaborators (see Babuška and Suri (1994)). While the theory for this type of approximation is far from complete, it will be useful to mention a few facts that separate it from the free knot case discussed above.

Let Σ^*_n denote the set of all piecewise polynomials

$$S = \sum_{I\in\Delta} P_I \chi_I, \tag{6.24}$$

where Δ is a partition and for each $I \in \Delta$ there is a polynomial P_I of order r_I with $\sum_{I\in\Delta} r_I \le n$. As usual, we let

$$\sigma^*_n(f)_p := \inf_{S\in\Sigma^*_n} \|f - S_n\|_{L_p(\Omega)}. \tag{6.25}$$

Clearly, for each $r = 1, 2, \ldots$, we have $\sigma^*_{nr}(f)_p \le \sigma_{n,r}(f)_p$ because $\Sigma_{n,r} \subset \Sigma^*_{nr}$. To see that σ^*_n can be considerably better than $\sigma_{n,r}$, we consider the following example, which was studied in DeVore and Scherer (1980). Let $f(x) = x^\beta$ with $\beta > 0$. We have seen that $\sigma_{n,r}(f)_p \asymp n^{-r}$. On the other hand, it is shown in the above reference that

$$\sigma^*_n(f) \le Ce^{-c\sqrt{n}}, \quad c := \sqrt{2} - 1 \tag{6.26}$$

and that this estimate cannot be improved in the sense of a better exponential rate.

6.5. Free partition splines: the multivariate case

Up to this point our discussion of nonlinear approximation has been almost entirely limited to approximating univariate functions. The question arises, for example, whether the results of the previous section on free knot spline approximation can be extended to the multivariate case.

For the moment, we restrict our discussion to the bivariate case and approximation on $\Omega := [0,1]^2$. In this case, we consider the space $\Sigma_{n,r}^{\#}$ consisting of all functions

$$S = \sum_{T\in\Delta} P_T \chi_T \tag{6.27}$$

with $\Delta = \{T\}$ a partition of Ω consisting of n triangles and the P_T polynomials of total order r on T for each $T \in \Delta$. Let

$$\sigma_{n,r}^{\#}(f)_p := \inf_{S\in\Sigma_{n,r}^{\#}} \|f - S\|_{L_p(\Omega)}. \tag{6.28}$$

Here # is used to make a distinction from the univariate case.

There is no known characterization of $\mathcal{A}_q^\alpha(L_p(\Omega), (\Sigma_{n,r}^{\#}))$ for any values of α, p, q. This remains one of the most interesting and challenging problems in nonlinear approximation. We shall mention some of the difficulties encountered in trying to characterize these approximation classes, since this has influenced developments in multivariate nonlinear approximation.

A first remark is that the space $\Sigma_N^{\#}$ does not satisfy assumption (iv) of Section 4.1: that is, for no constant c do we have $\Sigma_n^{\#} + \Sigma_n^{\#} \subset \Sigma_{cn}^{\#}$. For instance, consider a partition Δ_1 of Ω consisting of n vertical strips of equal size, each divided into two triangles, and the corresponding partition Δ_2 made from horizontal strips. Let S_1 be a piecewise polynomial relative to Δ_1 and S_2 another piecewise polynomial relative to Δ_2. Then the sum $\Delta_1 + \Delta_2$ will be a piecewise polynomial which in general requires $4n^2$ triangles in its partition.

Even more relevant to our problem is a result (communicated to us by Jonathan Goodman) that constructs functions $f(x)$ and $g(y)$ which individually can be approximated with order $O(1/n)$ by the elements of $\Sigma_n^{\#}$ but whose sum can only be approximated to order $O(1/\sqrt{n})$. Thus, the approximation spaces $\mathcal{A}_q^\alpha(L_p(\Omega))$ are not linear. This precludes their characterization by classical smoothness spaces, which are always linear.

Here is another relevant comment. The starting point for proving direct estimates for nonlinear piecewise polynomial approximation are good local error estimates for polynomial approximation, such as those given in Section 6.1. The appropriate local error estimators for polynomial approximation on general triangles are not known. They should take into consideration the shape and orientation of the triangles. For example, less smoothness of the target function should be required in directions where the triangle is thin, more in directions where the triangle is fat. While one may guess appropriate error estimators, none have been utilized successfully in nonlinear schemes.

Given the situation described above concerning nonlinear piecewise polynomial approximation, it comes as no surprise that other avenues were explored to handle nonlinearity in the multivariate case. The most successful of these has been n-term approximation, which took the following viewpoint. In the univariate case the elements in the space Σ_n can also be desribed as a sum of n (or perhaps Cn) fundamental building blocks. In the case of piecewise constants these are simply the characteristic functions χ_I of intervals I. In the general case of univariate, nonlinear piecewise polynomial approximation the building blocks are B-splines. Therefore, one generalization of Σ_n to the multivariate case would take the form of n-term approximation using multivariate building blocks. The first examples were for box splines (DeVore and Popov 1987) but this was later abandoned for the more computationally favourable wavelets. We shall discuss wavelets in Section 7.

6.6. Rational approximation

Another natural candidate for nonlinear approximation is the set of rational functions. Let $\mathcal{R}_n(\mathbb{R}^d)$ denote the space of rational functions in d variables. Thus, an element R in $\mathcal{R}_n$ is the quotient, $R = P/Q$, of two polynomials P,Q (in d variables) of total degree $\leq n$. We define the approximation error

$$r_n(f)_p := \inf_{R\in\mathcal{R}_n} \|f - R\|_{L_p(\Omega)}. \tag{6.29}$$

The status of rational approximation is more or less the same as for piecewise polynomials. In one variable, we have

$$\mathcal{A}_q^\alpha(L_p(\Omega), (\mathcal{R}_n)) = \mathcal{A}_q^\alpha(L_p(\Omega), (\Sigma_{n,r})), \quad 0 < \alpha < r. \tag{6.30}$$

Thus, on the one hand the approximation problem is solved but on the other hand the news is somewhat depressing since there is nothing to gain or lose (in the context of the approximation classes) in choosing rational functions over piecewise polynomials.

The characterizations (6.30) were proved by Pekarski (1986) and Petrushev (1988) by comparing σ_n to r_n. A typical comparison is given by the inequalities

$$r_n(f)_p \leq Cn^{-\alpha} \sum_{k=1}^{n} k^{\alpha-1}\sigma_{k,r}(f)_p, \quad n \geq r, \tag{6.31}$$

which hold for all $1 \le p < \infty$, $\alpha > 0$ and approximation on an interval. Similar inequalities reverse the roles of $\sigma_{n,r}(f)_p$ and $r_n(f)_p$. Thus the approximation classes for univariate rational approximation coincide with Besov spaces when $1/q = \alpha + 1/p$ (see (6.23)). In a strong sense, rational approximation can be viewed as piecing together local polynomial approximants similar to piecewise polynomials.

We should also mention the work of Peller (1980), who characterized the approximation classes for rational approximation in the BMO metric (which can be considered as a slight variant of L_∞). In the process, Peller characterized interpolation spaces between BMO and the Besov space $B_1^1(L_1)$ and found the trace classes for Hankel operators, thus unifying three important areas of analysis.

There are some direct estimates for multivariate rational approximation (see, for example, DeVore and Yu (1990)) but they fall far short of being optimal. The characterization of approximation spaces for multivariate rationals has met the same resistance as piecewise polynomials for more or less the same reasons.

There have been several other important developments in rational approximation. One of these was Newman's theorem (see Newman (1964)) which showed that the function $f(x) = |x|$ satisfies $r_n(f)_\infty = O(e^{-c\sqrt{n}})$ (a very stunning result at the time). Subsequently, similar results were proved for other special functions (such as $e^{-|x|^\beta}$) and even asymptotics for the error $r_n(f)$ were found. A mainstay technique in these developments was Padé approximation. This is to rational functions what Taylor expansions are to polynomials. A first reference for Padé approximation is the book of Baker (1975).

7. Wavelets

Wavelets were ripe for discovery in the 1980s. Multigrid methods in numerical computation, box splines in approximation theory, and the Littlewood–Paley theory in harmonic analysis all pointed to multilevel decompositions. However, the great impetus came from two discoveries: the multiresolution analysis of Mallat and Meyer (see Mallat (1989)) and most of all the discovery by Daubechies (1988) of compactly supported orthogonal wavelets with arbitrary smoothness.

Wavelets are tailor-made for nonlinear approximation and certain numerical applications. Computation is fast and simple, and strategies for generating good nonlinear approximations are transparent. Since wavelets provide unconditional bases for a myriad of function spaces and smoothness spaces, the characterization of approximation classes is greatly simplified. Moreover, wavelets generalize readily to several dimensions.

There are many excellent accounts of multiresolution and wavelet theory. We shall introduce only enough of the theory to set our notation and provide us with the vehicle we need for our development of nonlinear approximation. The Haar function is a wavelet (albeit not a very smooth one) and (3.47) is typical of wavelet decompositions. We shall begin our discussion of multiresolution by considering approximation from shift invariant spaces which provides the *linear theory* for wavelet approximation.

In the development of wavelets and multiresolution analysis, one needs to make modest assumptions on the refinement function φ so the theory develops smoothly. We shall not stress these assumptions, and in fact in many cases not even mention them, in order to keep our exposition short and to the point. The reader needs to consult one of the following references to find precise formulations of the results we state here: Daubechies (1992), Meyer (1990), DeVore and Lucier (1992).

7.1. Shift invariant spaces

In multiresolution analysis, there are two fundamental operations we perform on functions: shift and dilation. If f is defined on $\mathbb{R}^d$ and $j \in \mathbb{Z}^d$, then $f(\cdot - j)$ is the (integer) shift of f by j. Meanwhile, if $a > 0$ is a real number then $f(a\cdot)$ is the dilate of f by a. In this section, we consider spaces invariant under shifts. We then dilate them to create new and finer spaces. The main goal is to understand the approximation properties of these dilated spaces.

We shall not discuss *shift invariant spaces* in their full generality in order to move more directly to multiresolution analysis. The results stated below have many extensions and generalizations (see de Boor, DeVore and Ron (1993) and the references therein).

Let φ be a compactly supported function in $L_2(\mathbb{R}^d)$. We define $\tilde{\mathcal{S}}(\varphi)$ as the set of all finite linear combinations of the shifts of φ. The space $\mathcal{S} := \mathcal{S}(\varphi)$ is defined to be the closure of $\tilde{\mathcal{S}}(\varphi)$ in $L_2(\mathbb{R}^d)$. We say that $\mathcal{S}$ is the *principal shift invariant* (PSI) *space* generated by φ.

For each $k \geq 0$, the space $\mathcal{S}_k := \mathcal{S}_k(\varphi)$ is defined to be the dilate of $\mathcal{S}$ by 2^k. A function T is in $\mathcal{S}_k$ if and only if $T = S(2^k\cdot)$ with $S \in \mathcal{S}(\varphi)$. The space $\mathcal{S}_k$ is invariant under the shifts $j2^{-k}$, $j \in \mathbb{Z}^d$. We shall be interested in the approximation properties (in the $L_2(\mathbb{R}^d)$-norm) of $\mathcal{S}_k$ as $k \to \infty$. We let

$$E_k(f) := E_k(f)_2 := \inf_{S \in \mathcal{S}_k} \|f - S\|_{L_2(\mathbb{R}^d)}, \quad k = 0, 1, \ldots. \tag{7.1}$$

The approximation properties of $\mathcal{S}_k$ are related to polynomial reproduction in $\mathcal{S}$. It was Schoenberg (1946) who first recognized that polynomial reproduction could be described by the Fourier transform $\hat{\varphi}$ of φ; subsequently, Strang and Fix (1973) used Fourier transforms to describe approximation properties. We say φ satisfies the Strang–Fix condition of order $r \in \mathbb{N}$ if

$$\hat{\varphi}(0) \neq 0, \text{ and } D^j\hat{\varphi}(2k\pi) = 0, \quad k \in \mathbb{Z}^d \setminus \{0\}, |j| < r. \tag{7.2}$$

When φ satisfies the Strang–Fix condition of order r then $\mathcal{S}(\phi)$ locally contains all polynomials of order r (degree $< r$). (Actually, this and results stated below require a little more about φ in terms of smoothness, which we choose not to formulate exactly.) Moreover, it is easy to prove the Jackson inequality: for all f in the Sobolev space $W^r(L_2(\mathbb{R}^d))$, we have

$$E_k(f) \leq C2^{-kr}|f|_{W^r(L_2(\mathbb{R}^d))}, \quad k = 0, 1, \ldots. \tag{7.3}$$

The companion Bernstein inequality to (7.3) is

$$|S|_{W^r(L_2(\mathbb{R}^d))} \leq C2^{kr}\|S\|_{L_2(\mathbb{R}^d)}, \quad S \in \mathcal{S}_k. \tag{7.4}$$

It is valid if φ is in $W^r(L_2(\mathbb{R}^d))$. Under these conditions on φ, we can use the general results of Section 4.3 to obtain the following characterization of approximation spaces:

$$\mathcal{A}_q^\alpha(L_2(\mathbb{R}^d)) = B_q^\alpha(L_2(\mathbb{R}^d)), \quad 0 < \alpha < r, \ 0 < q \leq \infty. \tag{7.5}$$

Notice that this is exactly the same characterization as for the other types of linear approximation we have discussed earlier. There is a similar theory for approximation in $L_p(\mathbb{R}^d)$, $1 \leq p \leq \infty$, and even $0 < p < 1$.

7.2. Multiresolution and wavelet decompositions

Multiresolution adds one essential new ingredient to the setting of the previous section. We require that the spaces $\mathcal{S}_k$ are nested, that is, $\mathcal{S}_k \subset \mathcal{S}_{k+1}$, which is of course equivalent to $\mathcal{S}_0 \subset \mathcal{S}_1$. This is in turn equivalent to requiring that φ is in $\mathcal{S}_1$.

We shall limit our discussion to the multiresolution analysis that leads to the biorthogonal wavelets of Cohen, Daubechies and Feauveau (1992). These are the wavelets used most often in applications. Accordingly, we start with the univariate case and assume that φ is a function for which the spaces $\mathcal{S}_k = \mathcal{S}_k(\varphi)$ of the previous section provide approximation:

$$\operatorname{dist}(f, \mathcal{S}_k)_{L_2(\mathbb{R})} \to 0. \tag{7.6}$$

We know that this will hold, for example, if φ satisfies the Strang–Fix condition for some order $r > 0$. We assume further that the shifts $\varphi(\cdot - j)$, $j \in \mathbb{Z}$, are a Riesz basis for $\mathcal{S}$ and that the dual basis is given by the shifts of a compactly supported function $\tilde{\varphi}$ whose dilated spaces $\mathcal{S}_k(\tilde{\varphi})$ also form a multiresolution analysis. Duality means that

$$\int_{\mathbb{R}} \varphi(x - j)\tilde{\varphi}(x - k)\,\mathrm{d}x = \delta_{jk}. \tag{7.7}$$

with δ_{jk} the Kronecker delta.

The fact that $\varphi \in \mathcal{S}_1$ implies that φ is refinable:

$$\varphi(x) = \sum_{k\in\mathbb{Z}} c_k \varphi(2x-k). \tag{7.8}$$

The compact support of φ implies that there is only a finite number of nonzero coeffecents c_k in (7.8). They are called the refinement mask for φ (in image processing they are called the (low pass) filter coefficients). The dual function $\tilde{\varphi}$ satisfies a corresponding refinement equation with mask coefficients $\tilde{c}_k$.

Let $\langle\cdot,\cdot\rangle$ denote the inner product in $L_2(\mathbb{R})$ and let P be the projector

$$Pf := \sum_{j\in\mathbb{Z}} \langle f, \tilde{\varphi}(\cdot - j)\rangle \varphi(\cdot - j) \tag{7.9}$$

which maps $L_2(\mathbb{R})$ onto $\mathcal{S}$. By dilation, we obtain the corresponding projectors P_k which map $L_2(\mathbb{R})$ onto $\mathcal{S}_k$, $k \in \mathbb{Z}$. We are particularly interested in the projector $Q := P_1 - P_0$ which maps $L_2(\mathbb{R})$ onto a subspace W of $\mathcal{S}_1$. The space W is called a wavelet space; it represents the *detail* which, when added to $\mathcal{S}_0$, gives $\mathcal{S}_1$ via the formula $S = PS + QS$, $S \in \mathcal{S}_1$. One of the main results of wavelet/multiresolution theory is that W is a PSI space generated by the function

$$\psi(x) = \sum_{k\in\mathbb{Z}} d_k \tilde{\varphi}(2x-k), \quad d_k := (-1)^k \tilde{c}_{1-k}. \tag{7.10}$$

Also, the shifts $\psi(\cdot - j)$, $j \in \mathbb{Z}$, form a Riesz basis for W whose dual functionals are represented by $\tilde{\psi}(\cdot - j)$ where $\tilde{\psi}$ is obtained from φ in the same way ψ was obtained from $\tilde{\varphi}$. In other words,

$$Qf = \sum_{j\in\mathbb{Z}} 2^{-k} \langle f, \tilde{\psi}(\cdot - j)\rangle \psi(\cdot - j). \tag{7.11}$$

Of course, by dilation, we obtain the spaces W_k, the projectors Q_k and the representation

$$Q_k f = \sum_{j\in\mathbb{Z}} 2^k \langle f, \tilde{\psi}(2^k \cdot - j)\rangle \psi(2^k \cdot - j). \tag{7.12}$$

From (7.6), we know that $P_k f \to f$, $k \to \infty$. It can also be shown that $P_k f \to 0$, $k \to -\infty$, and therefore we have

$$f = \sum_{k=-\infty}^{\infty} (P_{k+1}f - P_k f) = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} 2^k \langle f, \tilde{\psi}(2^k \cdot - j)\rangle \psi(2^k \cdot - j). \tag{7.13}$$

The factor 2^k multiplying the inner product arises from scaling. This is the biorthogonal wavelet decomposition of an arbitrary $f \in L_2(\mathbb{R})$. We would like to simplify the wavelet notation and better expose the nature of the representation (7.13). For this we shall use the following convention. To $j \in \mathbb{Z}^d$, $k \in \mathbb{Z}$, we associate the dyadic cube $I = 2^{-k}(j+\Omega)$ with $\Omega := [0,1]^d$, the unit cube in $\mathbb{R}^d$. To each function η defined on $\mathbb{R}^d$, we let

$$\eta_I(x) := |I|^{-1/2}\eta(2^k \cdot -j). \tag{7.14}$$

The cube I roughly represents the support of η_I; in the case that $\eta = \chi_\Omega$ or $\eta = H$ with the H the Haar function, then I is precisely the support of η_I.

Let D be the set of all dyadic intervals in $\mathbb{R}$ and D_k those dyadic intervals of length 2^{-k}. We can now rewrite (7.13) as

$$f = \sum_{I \in D} c_I(f)\psi_I, \quad c_I(f) := \langle f, \psi_I \rangle. \tag{7.15}$$

The Riesz basis property of the ψ_I gives that

$$\|f\|_{L_2(\mathbb{R})} \asymp \left(\sum_{I \in D} |c_I(f)|^2 \right)^{1/2}. \tag{7.16}$$

The special case of orthogonal wavelets is noteworthy. In this case, one begins with a scaling function φ whose shifts are an orthonormal system for $\mathcal{S}(\varphi)$. Thus $\tilde{\varphi} = \varphi$ and the space W is orthogonal to $\mathcal{S}_0$: each function $S \in W$ satisfies

$$\int_{\mathbb{R}} S S_0 \, \mathrm{d}x = 0, \quad S_0 \in \mathcal{S}_0. \tag{7.17}$$

The decomposition $\mathcal{S}_1 = \mathcal{S}_0 \oplus W$ is orthogonal and the functions ψ_I, $I \in D$ are an orthonormal basis for $L_2(\mathbb{R})$.

We turn now to the construction of wavelet bases in several dimensions. There are several possibilities. The most often used construction is the following. Let φ be a univariate scaling function and ψ its corresponding wavelet. We define $\psi^0 := \varphi$, $\psi^1 := \psi$. Let E' denote the collection of vertices of the unit cube $[0,1]^d$ and E the set of nonzero vertices. For each vertex $e = (e_1, \ldots, e_d) \in E'$, we define the multivariate functions

$$\psi^e(x_1, \ldots, x_d) := \psi^{e_1}(x_1) \cdots \psi^{e_d}(x_d) \tag{7.18}$$

and define $\Psi := \{\psi^e : e \in E\}$. If $D = D(\mathbb{R}^d)$ is the set of dyadic cubes in $\mathbb{R}^d$, then the collection of functions

$$\{\psi_I^e, \quad I \in D, \ e \in E\} \tag{7.19}$$

forms a Riesz basis for $L_2(\mathbb{R}^d)$; an orthonormal basis if ψ is an orthogonal wavelet. The dual basis functions $\tilde{\psi}_I^e$ have an identical construction starting with $\tilde{\varphi}$ and $\tilde{\psi}$. Thus, each $f \in L_2(\mathbb{R}^d)$ has the wavelet expansion

$$f = \sum_{I \in D} \sum_{e \in E} c_I^e(f)\psi_I^e, \quad c_I^e(f) := \langle f, \tilde{\psi}_I^e \rangle. \tag{7.20}$$

Another construction of multivariate wavelet bases is to simply take the tensor products of the univariate basis ψ_I. This gives the basis

$$\psi_R(x_1,\dots,x_d) := \psi_{I_1}(x_1)\cdots\psi_{I_d}(x_d), \quad R := I_1\times\cdots I_d, \tag{7.21}$$

where the R are multidimensional parallelepipeds. Notice that the support of the function ψ_R corresponds to R and is nonisotropic. It can be long in one direction and short in another. This is in contrast to the previous bases whose supports are isotropic. We shall be almost exclusively interested in the first basis.

7.3. Characterization of function spaces by wavelet coefficients

Wavelet coefficients provide simple characterizations of most function spaces. The norm in the function space is equivalent to a sequence norm applied to the wavelet coefficients. We shall need such characterizations for the case of L_p spaces and Besov spaces.

It is sometimes convenient in the characterizations that follow to choose different normalizations for the wavelets, and hence coefficients, appearing in the decomposition (7.20). In (7.20) we have normalized the wavelets and dual functions in $L_2(\mathbb{R}^d)$. We can also normalize the wavelets in $L_p(\mathbb{R}^d)$, $0<p\le\infty$, by taking

$$\psi^e_{I,p} := |I|^{-1/p+1/2}\psi^e_I, \quad I\in D,\ e\in E, \tag{7.22}$$

with a similar definition for the dual functions. Then, we can rewrite (7.20) as

$$f=\sum_{I\in D}\sum_{e\in E} c^e_{I,p}(f)\psi^e_{I,p}, \quad c^e_{I,p}(f) := \langle f,\tilde\psi^e_{I,p'}\rangle, \tag{7.23}$$

with $1/p+1/p'=1$. We also define

$$c_{I,p}(f) := \left(\sum_{e\in E}|c^e_{I,p}(f)|^p\right)^{1/p}. \tag{7.24}$$

One should note that it is easy to go from one normalization to another. For example, for any $0,p,q\le\infty$, we have

$$\psi_{I,p} = |I|^{1/q-1/p}\psi_{I,q}, \quad c_{I,p}(f) = |I|^{1/p-1/q}c_{I,q}(f). \tag{7.25}$$

The characterization of L_p spaces by wavelet coefficients comes from the Littlewood–Paley theorem of harmonic analysis. One cannot simply characterize the L_p spaces by ℓ_p norms of the wavelet coefficients. Rather, one must go through the square function

$$S(f,x) := \left(\sum_{I\in D} c_{I,2}(f)^2|I|^{-1}\chi_I(x)\right)^{1/2} = \left(\sum_{I\in D} c_{I,p}(f)^2|I|^{-2/p}\chi_I(x)\right)^{1/2} \tag{7.26}$$

which incorporates the interaction between dyadic levels. Here, as earlier, χ_I is the characteristic function of the interval I. For $1 < p < \infty$, one has

$$\|f\|_{L_p(\mathbb{R}^d)} \asymp \|S(f,\cdot)\|_{L_p(\mathbb{R})} \tag{7.27}$$

with the constants of equivalency depending only on p. Notice that, in the case $p = 2$, (7.27) reduces to (7.16). One can find proofs of (7.27) (which use techniques of harmonic analysis such as maximal functions) in Meyer (1990) or DeVore, Konyagin and Temlyakov (1998).

The equivalence (7.27) can be extended to the range $p \le 1$ if the space L_p is replaced by the Hardy space H_p and more assumptions are made of the wavelet ψ. In this sense, most of the theory of approximation given below can be extended to this range of p.

We have introduced the Besov spaces $B_q^\alpha(L_p(\mathbb{R}^d))$ for $0 < q, p \le \infty$, $\alpha > 0$, in Section 4.5. The following is the wavelet characterization of these spaces:

$$|f|_{B_q^\alpha(L_p(\mathbb{R}^d))} \asymp \begin{cases} \left(\sum_{k=-\infty}^{\infty} 2^{k\alpha q}\left(\sum_{I\in D_k} c_{I,p}(f)^p\right)^{q/p}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{k\in\mathbb{Z}} 2^{k\alpha}\left(\sum_{I\in D_k} c_{I,p}(f)^p\right)^{1/p}, & q = \infty. \end{cases} \tag{7.28}$$

Several remarks are in order to explain (7.28).

Remark 7.1 Other normalizations for the coefficients $c_I(f)$ are frequently used. The form of (7.28) then changes by the introduction of a factor $|I|^\beta$ into each term, with β a fixed constant.

Remark 7.2 We can define spaces of functions for all $\alpha > 0$ by using the right side of (7.28). However, these spaces will coincide with Besov spaces only for a certain range of α and p that depend on the wavelet ψ. In the case $1 \le p \le \infty$, we need that

(a) $\psi \in B_q^\beta(L_p(\mathbb{R}^d))$, for some $\beta > \alpha$,
(b) ψ has r vanishing moments with $r > \alpha$.

When $p < 1$, we also need that $r > d/p - d$ (see the following remark).

Remark 7.3 When $p < 1$, (7.28) characterizes the space $B_q^\alpha(H_p(\mathbb{R}^d))$ (with the correct range of parameters) where this latter Besov space can be defined by replacing the L_p modulus of smoothness by the H_p modulus of smoothness (see Kyriazis (1996)). However, if $\alpha > d/p - d$, this space is the same as $B_q^\alpha(L_p(\mathbb{R}^d))$.

Remark 7.4 For a fixed value of $1 \leq p < \infty$, the spaces $B_\tau^\alpha(L_\tau(\mathbb{R}^d))$, $1/\tau = \alpha/d + 1/p$, occur, as we know, in nonlinear approximation. If we choose the wavelets normalized in L_p, then the characterization (7.28) becomes simply

$$|f|_{B_\tau^\alpha(L_\tau(\mathbb{R}^d))} \asymp \left(\sum_{I \in D} c_{I,p}(f)^\tau \right)^{1/\tau}. \tag{7.29}$$

7.4. Nonlinear wavelet approximation

In this and the next subsections, we shall consider n-term approximation by wavelet sums. The results we present hold equally well in the univariate and the multivariate case. However, the notation is somewhat simpler in the univariate case. Therefore, to spare the reader, we shall initially treat only this case. At the end of the section we shall formulate the results for multivariate functions.

The idea of how to utilize wavelets in nonlinear approximation is quite intuitive. If the target function is smooth on a region we can use a coarse resolution (approximation) on that region. This amounts to putting terms in the approximation corresponding to low frequency-terms from dyadic level k with k small. On regions where the target function is not smooth we use higher resolution. This is accomplished by taking more wavelet functions in the approximation, that is, terms from higher dyadic levels. The questions that arise from these intuitive observations are:

(i) exactly how should we measure smoothness to make such demarcations between smooth and nonsmooth?
(ii) how do we allocate terms in a nonlinear strategy?
(iii) are there precise characterizations of the functions that can be approximated with a given approximation order by nonlinear wavelet approximation?

Fortunately, all of these questions have a simple and definitive solution, which we shall presently describe.

We shall limit ourselves to the case of biorthogonal wavelets and approximation in L_p, $1 < p < \infty$. Again, one can work in much more generality. As will be clear from our exposition, what is essential is only the equivalence of function norms with norms on the sequence of wavelet coefficients. Thus, the results we present hold equally well for approximation in the Hardy space H_p (Cohen, DeVore and Hochmuth 1997) and for more general wavelets.

It will also be convenient to consider approximation on all of $\mathbb{R}^d$ (initially on $\mathbb{R}$). In the following section, we shall discuss briefly how results extend to other domains.

Let φ, $\tilde{\varphi}$ be two refinable functions which are in duality as described in

Section 7.2 and let ψ and $\tilde{\psi}$ be their corresponding wavelets. Then, each function $f \in L_p(\mathbb{R})$ has the wavelet decomposition (7.15). We let Σ_n^w denote the set of all functions

$$S = \sum_{I \in \Lambda} a_I \psi_I, \tag{7.30}$$

where $\Lambda \subset D$ is a set of dyadic intervals of cardinality $\#\Lambda \leq n$. Thus Σ_n^w is the set of all functions which are a linear combination of n wavelet functions. In analogy with our previous studies, we define

$$\sigma_n^w(f)_p := \inf_{S \in \Sigma_n^w} \|f - S\|_{L_p(\mathbb{R})}. \tag{7.31}$$

We can characterize the approximation classes for n-term wavelet approximation by proving Jackson and Bernstein inequalities and then invoking the general theory of Section 4.3. The original proofs of these inequalities were given in DeVore, Jawerth and Popov (1992) but we shall follow Cohen, DeVore and Hochmuth (1997) which introduced some simpler techniques.

Given a finite set Λ of intervals, for each $x \in \mathbb{R}$, we let $I(x)$ be the smallest interval in Λ that contains x. If there is no such interval, then we define $I(x) := \mathbb{R}$ and expressions like $|I(x)|^{-1}$ are interpreted as zero. The following lemma of Temlyakov (1998*a*) is a powerful tool in estimating norms of wavelet sums.

Lemma 1 Let $1 < p < \infty$ and Λ be a finite set. If $f \in L_p(\mathbb{R})$ has the wavelet decomposition

$$f = \sum_{I \in \Lambda} c_{I,p}(f) \psi_{I,p}, \tag{7.32}$$

with $|c_{I,p}(f)| \leq M$, for all $I \in \Lambda$, then

$$\|f\|_{L_p(\mathbb{R})} \leq C_1 M \#\Lambda^{1/p}, \tag{7.33}$$

with C_1 an absolute constant. Similarly, if $|c_{I,p}(f)| \geq M$, for all $I \in \Lambda$, then

$$\|f\|_{L_p(\mathbb{R})} \geq C_2 M \#\Lambda^{1/p}, \tag{7.34}$$

with $C_2 > 0$ an absolute constant.

We shall sketch the proof of (7.33) (which is valid for $0 < p < \infty$) since it gives us a chance to show the role of $I(x)$ and the square function. The proof of (7.34) is similar. We have

$$\begin{aligned} \|f\|_{L_p(\mathbb{R})} &\leq \|S(f)\|_{L_p(\mathbb{R})} = C \left\| \left(\sum_{I \in \Lambda} c_{I,p}^2 |I|^{-2/p} \chi_I \right)^{1/2} \right\|_{L_p(\mathbb{R})} \\ &\leq CM \left\| \left(\sum_{I \in \Lambda} |I|^{-2/p} \chi_I \right)^{1/2} \right\|_{L_p(\mathbb{R})} \leq CM \| |I(x)|^{-1/p} \|_{L_p(\mathbb{R})}. \end{aligned}$$

If $J \in \Lambda$, then the set $\tilde{J} := \{x : I(x) = J\}$ is a subset of J. It follows that

$$\|f\|_{L_p(\mathbb{R})}^p \le CM^p \int_{\mathbb{R}^d} |I(x)|^{-1}\,\mathrm{d}x \le CM^p \sum_{J\in\Lambda} \int_{\tilde{J}} |J|^{-1} \le CM^p \#\Lambda,$$

which proves (7.33).

We shall now formulate the Jackson inequality for n-term wavelet approximation. Let r be the number of vanishing moments of ψ. Recall that r also represents the order of polynomials that are locally reproduced in $\mathcal{S}(\varphi)$. Recall also that, for $0 < \tau < \infty$, a sequence (a_n) of real numbers is in the Lorentz space $w\ell_\tau := \ell_{\tau,\infty}$ if and only if

$$\#\{n : |a_n| > \epsilon\} \le M^\tau \epsilon^{-\tau} \tag{7.35}$$

for all $\epsilon > 0$. The norm $\|(a_n)\|_{w\ell_\tau}$ is the smallest value of M such that (7.35) holds. Also,

$$\|(a_n)\|_{w\ell_\tau} \le \|(a_n)\|_{\ell_\tau}.$$

Theorem 5 Let $1 < p < \infty$, and $s > 0$, and let $f \in L_p(\mathbb{R})$ and $c_I := c_{I,p}(f)$, $I \in D$, be such that $(c_I)_{I\in D}$ is in $w\ell_\tau$, $1/\tau = s + 1/p$. Then,

$$\sigma_n(f)_p \le Cn^{-s}\|(c_I)\|_{w\ell_\tau}, \quad n = 1, 2, \ldots, \tag{7.36}$$

with the constant C depending only on p and s.

We sketch the proof. We have

$$\#\{I : |c_I| > \epsilon\} \le M^\tau \epsilon^{-\tau}$$

for all $\epsilon > 0$ with $M := \|(c_I)\|_{w\ell_\tau}$. Let $\Lambda_j := \{I : 2^{-j} < |c_I| \le 2^{-j+1}\}$. Then, for each $k = 1, 2, \ldots$, we have

$$\sum_{j=-\infty}^{k} \#\Lambda_j \le CM^\tau 2^{k\tau} \tag{7.37}$$

with C depending only on τ.

Let $S_j := \sum_{I\in\Lambda_j} c_I\psi_I$ and $T_k := \sum_{j=-\infty}^{k} S_j$. Then $T_k \in \Sigma_N$ with $N = CM^\tau 2^{k\tau}$. We have

$$\|f - T_k\|_{L_p(\mathbb{R})} \le \sum_{j=k+1}^{\infty} \|S_j\|_{L_p(\mathbb{R})}. \tag{7.38}$$

We fix $j > k$ and estimate $\|S_j\|_{L_p(\mathbb{R})}$. Since $|c_I| \le 2^{-j+1}$ for all $I \in \Lambda_j$, we have, from Lemma 1 and (7.37),

$$\|S_j\|_{L_p(\mathbb{R})} \le C2^{-j}\#\Lambda_j^{1/p} \le CM^{\tau/p}2^{j(\tau/p-1)}.$$

We therefore conclude from (7.38) that

$$\|f - T_k\|_{L_p(\mathbb{R})} \le CM^{\tau/p} \sum_{j=k+1}^{\infty} 2^{j(\tau/p-1)} \le CM(M2^k)^{\tau/p-1}$$

because $\tau/p - 1 < 0$. In other words, for $N \asymp M^\tau 2^{k\tau}$, we have

$$\sigma_N(f)_p \le CMN^{1/p-1/\tau} = CMN^{-s}.$$

From the monotonicity of σ_n it follows that the last inequality holds for all $N \ge 1$.

Let us note a couple of things about the theorem. First of all there is no restriction on s. However, for large s, the set of functions satisfying $(c_{I,p}(f)) \in w\ell_\tau$ is not a classical smoothness space. We can use the theorem to obtain Jackson inequalities in terms of Besov spaces by using the characterization of Besov spaces by wavelet coefficients. Recall that this characterization applies to $B^s_\tau(L_\tau(\mathbb{R}))$ provided the following two properties hold:

(i) ψ has r vanishing moments with $r > s$
(ii) ψ is in $B^\rho_q(L_\tau)$ for some q and some $\rho > s$.

That is, ψ must have sufficient vanishing moments and sufficient smoothness. Under these assumptions, we have the following result.

Corollary 1 Let $1 < p < \infty$, let $s > 0$ and let $f \in B^s_\tau(L_\tau(\mathbb{R}))$, $1/\tau = s + 1/p$. If ψ satisfies the above two conditions (i) and (ii), then

$$\sigma_n(f)_p \le C|f|_{B^s_\tau(L_\tau(\mathbb{R}))} n^{-s}, \quad n = 1, 2, \ldots, \tag{7.39}$$

with C depending only on p and s.

We have $c_{I,\tau}(f) = c_{I,p}(f)|I|^{1/\tau-1/p} = c_{I,p}(f)|I|^{s/d}$. Thus, from (7.29) we find

$$|f|_{B^s_\tau(L_\tau(\mathbb{R}))} = \|(c_I)\|_{\ell_\tau} \ge \|(c_I)\|_{w\ell_\tau}.$$

Hence (7.39) follows from Theorem 5.

7.5. The Bernstein inequality for n-term wavelet approximation

The following theorem gives the Bernstein inequality which is the companion to (7.39).

Theorem 6 Let $1 < p < \infty$, and let the assumptions of Theorem 5 be valid. If $f = \sum_{I \in \Lambda} c_{I,p}(f)\psi_{I,p}$ with $\#\Lambda \le n$, we have

$$\|f\|_{B^s_\tau(L_\tau(\mathbb{R}))} \le Cn^s \|f\|_{L_p(\mathbb{R})}. \tag{7.40}$$

We sketch the simple proof of this inequality. We first note that, for each $I \in \Lambda$, we have

$$c_I |I|^{-1/p} \chi_I \le S(f),$$

because the left side is one of the terms appearing in the square function $S(f)$. Hence, with $I(x)$ defined as the smallest interval in Λ that contains x, we have, from (7.29),

$$\begin{aligned}
|f|^\tau_{B^s_\tau(L_\tau(\mathbb{R}))} &= \int_{\mathbb{R}} \sum_{I\in\Lambda} |c_I|^\tau |I|^{-1} \chi_I = \int_{\mathbb{R}} \sum_{I\in\Lambda} c_I^\tau |I|^{-\tau/p} \chi_I |I|^{-1+\tau/p} \chi_I \\
&\le C \int_{\mathbb{R}} S(f)^\tau \sum_{I\in\Lambda} |I|^{-1+\tau/p} \chi_r \le C \int_{\mathbb{R}} S(f,x)^\tau |I(x)|^{-1+\tau/p} \,\mathrm{d}x \\
&\le C \left(\int_{\mathbb{R}} S(f,x)^p \right)^{\tau/p} \left(\int_{\mathbb{R}} |I(x)|^{-1} \right)^{1-\tau/p} \mathrm{d}x \\
&\le Cn^{1-\tau/p} \|S(f)\|^\tau_{L_p(\mathbb{R})} \le Cn^{1-\tau/p} \|f\|^\tau_{L_p(\mathbb{R})}.
\end{aligned}$$

7.6. *Approximation spaces for n-term wavelet approximation*

The Jackson and Bernstein inequalities of the previous sections are equally valid in $\mathbb{R}^d$. The only distinction is that $n^{\pm s}$ should be replaced by $n^{\pm s/d}$. The proofs are identical to the univariate case except for the more elaborate notation needed in the multivariate formulation.

With the Jackson and Bernstein inequalities in hand, we can apply the general machinery of Section 4.3 to obtain the following characterization of the approximation spaces for n-term wavelet approximation. We formulate the results for the multivariate case.

Let $1 < p < \infty$ and $s > 0$ and let $1/\tau := s/d + 1/p$. If ψ satisfies the vanishing moments and smoothness assumptions needed for the Jackson and Bernstein inequalities, then, for any $0 < \gamma < s$ and any $0 < q \le \infty$,

$$\mathcal{A}^{\gamma/d}_q(L_p(\mathbb{R}^d)) = (L_p(\mathbb{R}^d), B^s_\tau(L_\tau(\mathbb{R}^d)))_{\gamma/s,q}. \tag{7.41}$$

Several remarks are in order about (7.41).

Remark 7.5 We have seen the interpolation spaces on the right side of (7.41) before for free knot spline approximation and $d = 1$.

Remark 7.6 For each γ there is one value of q where the right side is a Besov space; namely, when $1/q = \gamma/d + 1/p$, the right side of (7.41) is the Besov space $B^\gamma_q(L_q(\mathbb{R}^d))$ with equivalent norms.

Remark 7.7 There is a description of the interpolation spaces on the right of (7.41) in terms of wavelet coefficients. Namely, a function is in the space $(L_p(\mathbb{R}^d), B^s_\tau(L_\tau(\mathbb{R}^d)))_{\gamma/s,q}$ if and only if $(c_{I,p}(f))_{I\in D}$ is in the Lorentz space

$\ell_{\mu,q}$ where $1/\mu := \gamma/d + 1/p$ and, in fact, we have

$$|f|_{\mathcal{A}_q^{\gamma/d}(L_p)} \asymp \|(c_{I,p}(f))\|_{\ell_{\mu,q}}.$$

This verifies Remark 7.6 that, in the case that $q = \mu$, then $\mathcal{A}_\mu^{\gamma/d}(L_p(\mathbb{R}^d)) = B_\mu^\gamma(\mathbb{R}^d))$ with equivalent norms.

These results can be proved by a slightly finer analysis of n-term wavelet approximation (see Cohen, DeVore and Hochmuth (1997) and Temlyakov (1998*a*))

There is a further connection between n-term approximation and interpolation that we wish to bring out. Let p, s, and τ have the same meaning as above. For each n, let f_n denote a best n-term approximation to f in $L_p(\mathbb{R}^d)$ (which can be shown to exist – see Temlyakov (1998*a*)). It follows from what we have proved and Theorem 3 of Section 4.3 that, for $n = 1, 2, \ldots$, we have

$$K(f, n^{-s}, L_p(\mathbb{R}^d), B_\tau^s(L_\tau(\mathbb{R}^d))) = \|f - f_n\|_{L_p(\mathbb{R}^d)} + n^{-s}|f_n|_{B_\tau^s(L_\tau(\mathbb{R}^d))}.$$

In other words, f_n realizes this K-functional at $t = n^{-s}$.

In summary, n-term wavelet approximation offers an attractive alternative to free knot spline approximation on several counts. In one space dimension (the only case where free knot spline approximation is completely understood), it provides the same approximation efficiency and yet is more easily numerically implementable (as will be discussed subsequently).

7.7. Wavelet decompositions and n-term approximation on domains in $\mathbb{R}^d$

In numerical considerations, we usually deal with functions defined on a finite domain $\Omega \subset \mathbb{R}^d$. The above results can be generalized to that setting in the following way. We assume that the boundary $\partial\Omega$ of of Ω is Lipschitz (it is possible to work under slightly weaker assumptions). Under this assumption, it follows that any function f in the Besov space $B_q^\alpha(\Omega)$ can be extended to all of $\mathbb{R}^d$ in such a way that the extended function Ef satisfies

$$|Ef|_{B_q^\alpha(L_p(\mathbb{R}^d))} \leq C|f|_{B_q^\alpha(L_p(\Omega))}. \tag{7.42}$$

We refer the reader to DeVore and Sharpley (1984, 1993) for a discussion of such extensions. The extended function Ef has a wavelet decomposition (7.23) and the results of the previous section can be applied. The n-term approximation to Ef will provide the same order of approximation to f on Ω and one can delete in the approximant all terms corresponding to wavelets that are not active on Ω (that is, all wavelets whose support does not intersect Ω).

While the above remarks concerning extensions are completely satisfactory for theoretical considerations, they are not always easily implementable in numerical settings. Another approach which is applicable in certain settings is the construction of a wavelet basis for the domain Ω. This is particularly suitable in the case of an interval $\Omega \subset \mathbb{R}$. Biorthogonal wavelet bases can be constructed for an interval (see Cohen, Daubechies and Vial (1993)) and can easily be extended to parallelepipeds in $\mathbb{R}^d$ and even polyhedral domains (see Dahmen (1997) and the references therein).

7.8. Thresholding and other numerical considerations

We have thus far concerned ourselves mainly with the theoretical aspects of n-term wavelet approximation. We shall now discuss how this form of approximation is implemented in practice. We assume that approximation takes place on a domain $\Omega \subset \mathbb{R}^d$ which admits a biorthogonal basis as discussed in the previous section. For simplicity of notation, we assume that $d = 1$. We shall also assume that the wavelet decomposition of the target function f is finite and known to us. This provides a good match with certain applications such as image processing. When the wavelet decomposition is not finite, one usually assumes more about f that allows truncation of the wavelet series while retaining the desired level of numerical accuracy.

In the case of approximation in $L_2(\Omega)$, the best n-term approximation to a target function f is obtained by choosing the n terms in the wavelet series (7.20) of f for which the coefficients are largest. A similar strategy applies in the case of $L_p(\mathbb{R})$ approximation. Now, we write f in its wavelet expansion with respect to L_p normalized wavelets (see (7.23)) and choose the n-terms for which $|c_{I,p}(f)|$ is largest. The results of Section 7.4 show that this approximant will provide the Jackson estimates for n-term wavelet approximation. It is remarkable that this simple strategy also gives a near best approximant f_n to f. Temlyakov (1998*a*) has shown that

$$\|f - f_n\|_{L_p(\Omega)} \leq C\sigma_n(f)_p, \quad n = 1, 2, \ldots, \tag{7.43}$$

with a constant C independent of f and n.

In numerical implementation, one would like to avoid the expensive sorting inherent in the above description of n-term approximation. This can be done by employing the following strategy known as *thresholding*. We fix the $L_p(\Omega)$ space in which the approximation error is to be measured. Given a tolerance $\epsilon > 0$, we let $\Lambda_\epsilon(f)$ denote the set of all intervals I for which $|c_{I,p}(f)| > \epsilon$ and define the *hard thresholding* operator

$$T_\epsilon(f) := \sum_{I \in \Lambda_\epsilon(f)} c_I(f)\psi_I = \sum_{|c_I(f)| > \epsilon} c_I(f)\psi_I. \tag{7.44}$$

If the target function f is in weak ℓ_τ, with $1/\tau = s + 1/p$, then it follows from the definition of this space that

$$\#(\Lambda_\epsilon) \leq M^\tau \epsilon^{-\tau} \tag{7.45}$$

Table 1. *Thresholding values*

Threshold	Number of coefficients	Error
ϵ	$M^\tau \epsilon^{-\tau}$	$M^{\tau/p}\epsilon^{1-\tau/p}$
$M^{-1/(ps)}\eta^{1/(s\tau)}$	$M^{1/s}\eta^{-1/s}$	η
$MN^{-1/\tau}$	N	MN^{-s}

with M the weak ℓ_τ norm of the coefficients. Moreover, arguing as in the proof of Theorem 5, we obtain

$$\|f - T_\epsilon(f)\|_{L_p(\Omega)} \le CM^{\tau/p}\epsilon^{1-\tau/p}. \tag{7.46}$$

For example, if $\epsilon = MN^{-1/\tau}$, then $\#(\Lambda_\epsilon(f)) \le N$ and $\|f - T_\epsilon(f)\|_{L_p(\Omega)} \le CMN^{-s}$. In other words, thresholding provides the Jackson estimate. In this sense, thresholding provides the same approximation efficiency as n-term approximation.

Table 1 records the relationship between thresholding and n-term approximation. Here, $M = |f|_{\ell_{\tau,\infty}}$, ϵ is a thresholding tolerance, η is a prescribed error, and N is a prescribed number of coefficients.

For example, the second row of this table gives bounds on the thresholding parameter and the number of coefficients needed to achieve an error tolerance $\eta > 0$.

Hard thresholding has a certain instability in that coefficients just below the thresholding tolerance are set to zero and those just above are kept intact. This instability can be remedied by *soft thresholding*. Given $\epsilon > 0$, we define

$$s_\epsilon(x) := \begin{cases} 0, & |x| \le \epsilon, \\ 2(|x| - \epsilon)\operatorname{sign} x, & \epsilon \le |x| \le 2\epsilon, \\ x, & |x| > 2\epsilon. \end{cases} \tag{7.47}$$

Then, the soft thresholding operator

$$T'_\epsilon(f) := \sum_{I \in D} s_\epsilon(c_{I,p}(f))\psi_{I,p} \tag{7.48}$$

has the same approximation properties as T_ϵ.

8. Highly nonlinear approximation

Nonlinear wavelet approximation in the form of n-term approximation or thresholding is simple and effective. However, two natural questions arise. How does the effectiveness of this form of approximation depend on the wavelet basis? Secondly, is there any advantage to be gained by adaptively choosing a basis which depends on the target function f? To be reasonable, we would have to limit our search of wavelet bases to a numerically implementable class. An example of such a class is the collection of wavelet packet bases defined in Section 3.7. We call such a class $\mathcal{L}$ of bases a *library*. We shall limit our discussion to approximation in a Hilbert space $\mathcal{H}$ and libraries of orthonormal bases for $\mathcal{H}$. So our problem of nonlinear approximation would be given a target function $f \in \mathcal{H}$, to choose both a basis $B \in \mathcal{L}$ and an n-term approximation to f from this basis. We call such an approximation problem *highly nonlinear* since it involves another layer of nonlinearity in the basis selection.

A closely related form of approximation is n-term approximation from a *dictionary* $\mathbb{D} \subset \mathcal{H}$ of functions. For us, a dictionary will be an arbitrary subset of $\mathcal{H}$. However, dictionaries have to be limited to be computationally feasible. Perhaps the first example of this type of approximation was considered by E. Schmidt (1907), who considered the approximation of functions $f(x,y)$ of two variables by bilinear forms $\sum_{i=1}^{m} u_i(x)v_i(y)$ in $L_2([0,1]^2)$. This problem is closely connected with properties of the integral operator with kernel $f(x,y)$.

We mention some other important examples of dictionaries. In neural networks, one approximates functions of d-variables by linear combinations of functions from the set

$$\{\sigma(a \cdot x + b) : a \in \mathbb{R}^d,\ b \in \mathbb{R}\},$$

where σ is a fixed univariate function. The functions $\sigma(a \cdot x + b)$ are planar waves; also called ridge functions. Usually, σ is required to have additional properties. For example, the sigmoidal functions, which are used in neural networks, are monotone nondecreasing, tend to 0 as $x \to -\infty$, and tend to 1 as $x \to \infty$.

Another example, from signal processing, uses the Gabor functions

$$g_{a,b}(x) := e^{iax} e^{-bx^2}$$

and approximates a univariate function by linear combinations of the elements from

$$\mathbb{D} := \{g_{a,b}(x - c) : a, b, c \in \mathbb{R}\}.$$

Gabor functions are one example of a dictionary of space(time)-frequency atoms. The parameter a serves to position the function $g_{a,b}$ in frequency and c does the same in space. The shape parameter b localizes $g_{a,b}$.

The common feature of these examples is that the family of functions used in the approximation process is redundant. There are many more functions in the dictionary than needed to approximate any target function f. The hope is that the redundancy will increase the efficiency of approximation.

On the other hand, redundancy may slow down the search for good approximations.

Results on highly nonlinear approximation are quite fragmentary and a cohesive theory still needs to be developed. We shall present some of what is known about this theory, both for its usefulness and in the hope of bringing attention to this interesting area.

8.1. Adaptive basis selection

It will be useful to begin by recalling the results of Sections 2 and 5 on n-term approximation using the elements of an orthonormal basis. Let $B := \{\eta_k\}$ be an orthonormal basis for $\mathcal{H}$ and let $\Sigma_n(B)$ denote the functions in $\mathcal{H}$ which can be written as a linear combination of n of the functions η_k, $k = 0, 1, \dots,$ and further let

$$\sigma_n(f, B) := \sigma_n(f, B)_{\mathcal{H}} := \inf_{S \in \Sigma_n(B)} \|f - S\|_{\mathcal{H}} \tag{8.1}$$

be the corresponding approximation error.

We have seen that the decrease of the approximation errors $\sigma_n(f, B)$ is completely determined by the rearranged coefficients $\langle f, \eta_k \rangle$. As before, we let $\gamma_k(f, B)$ be the kth largest of the absolute values of these coefficients. For example, we have seen that for any $\alpha > 0$, a function f from $\mathcal{H}$ is in $\mathcal{A}_\infty^\alpha$ (*i.e.*, $\sigma_n(f, B) = O(n^{-\alpha})$, $n \to \infty$), if and only if $(\gamma_n(f, B))$ is in weak ℓ_τ (*i.e.*, in $\ell_{\tau,\infty}$) with $\tau := (\alpha + 1/2)^{-1}$. Moreover,

$$\|(\gamma_n(f, B))\|_{\ell_{\tau,\infty}} \asymp |f|_{\mathcal{A}_\infty^\alpha}, \tag{8.2}$$

with constants of equivalency independent of B.

Suppose now that $\mathcal{L} = \{B\}$ is a library of such orthonormal bases B. We define the approximation error

$$\sigma_n^{\mathcal{L}}(f)_{\mathcal{H}} := \inf_{B \in \mathcal{L}} \sigma_n(f, B)_{\mathcal{H}}. \tag{8.3}$$

The approximation classes $\mathcal{A}_q^\alpha(\mathcal{H}, \mathcal{L})$ are defined in the usual way (see Section 4.1). It is of great interest to characterize the approximation classes in concrete settings since this would give us a clear indication of the advantages of adaptive basis selection. A few results are known in discrete settings (see, for instance, Kashin and Temlyakov (1997)). We shall limit ourselves to the following rather trivial observations.

In view of (8.2), we have the upper estimate

$$\sigma_n^{\mathcal{L}}(f)_{\mathcal{H}} \le C n^{-\alpha} \inf_B \|(\gamma_n(f, B))\|_{\ell_{\tau,\infty}} \tag{8.4}$$

with C an absolute constant. Moreover, for any $\alpha > 0$, we have

$$\cap_B \mathcal{A}_\infty^\alpha(\mathcal{H}, B) \subset \mathcal{A}_\infty^\alpha(\mathcal{H}, \mathcal{L}). \tag{8.5}$$

We can interpret these results in the following way. For each basis B, the condition $(\gamma_n(f)) \in \ell_{\tau,\infty}$, $\tau := (\alpha + 1/2)^{-1}$ can be viewed as a smoothness condition on f relative to the basis B. Thus the infimum on the right side of (8.4) can be thought of as the infimum of smoothness conditions relative to the different bases B. Similarly, we can view the classes $\mathcal{A}^\alpha_\infty(\mathcal{H}, B)$ as smoothness classes with respect to the basis B. The right side of (8.5) is an intersection of smoothness classes. Thus, an advantage of optimal basis selection is that we are allowed to take the basis $B \in \mathcal{L}$ in which f is smoothest.

In general (8.4) and (8.5) are not reversible. One can easily construct two basis B_1, B_2, and a target function f so that, as n varies, we alternate between the choices B_1 and B_2 as the best basis selection for varying n. It is less clear whether this remains the case when the library is chosen to have some structure as in the case of wavelet packets. Thus the jury is still out as to whether (8.4) and (8.5) can sometimes be reversed in concrete situations and thereby obtain a characterization of $\mathcal{A}^\alpha_\infty(\mathcal{H}, \mathcal{L})$.

The above discussion for $q = \infty$ generalizes to any $0 < q \le \infty$.

8.2. Two examples of wavelet libraries

We would be remiss in not mentioning at least a couple of simple examples of libraries of bases that are useful in applications. The understanding of the approximation properties in such examples would go a long way toward understanding highly nonlinear approximation.

Our first example is to generalize the wavelet packets of Section 3.7. Since the situation is completely analogous to that section, we shall be brief. In place of χ_Ω and the Haar function H, we can take any orthogonal scaling function φ and its orthogonal wavelet ψ. We take for $\mathcal{H}$ the space $L_2(\mathbb{R})$. The function φ satisfies the refinement equation (7.8) with refinement coefficients c_k, $k \in \mathbb{Z}$, and likewise the wavelet ψ satisfies (7.10). Therefore, the operators of (3.57) are replaced by

$$A_0 g := \sum_k c_k g(2 \cdot -k); \qquad A_1 g := \sum_k d_k g(2 \cdot -k). \tag{8.6}$$

Then, $A_0(\varphi) = \varphi$, and $A_1(\varphi) = \psi$.

Starting with $\gamma_0 := \varphi$, we generate the functions γ_b and the spaces Γ_b exactly as in Section 3.7. The interpretation using the binary tree of Figure 4 applies and gives the same interpretation of orthonormal bases for $\mathcal{S}_m(\varphi)$. These bases form the library of wavelet packet bases. For further discussion of wavelet packet libraries and their implementation, we refer the reader to Wickerhauser (1994).

For our second example, we take $\mathcal{H} = L_2(\mathbb{R}^2)$ and again consider a compactly supported, refinable function $\varphi \in L_2(\mathbb{R})$ with orthonormal shifts and its corresponding orthogonal wavelet ψ. We define $\psi^0 := \varphi$, $\psi^1 := \psi$. To each vertex e of the unit square $[0,1]^2$, each $j=(j_1,j_2)\in\mathbb{Z}^2$, $k=(k_1,k_2)\in\mathbb{Z}^2$, we associate the function

$$\psi^e_{j,k}(x_1,x_2):=2^{(k_1+k_2)/2}\psi^{e_1}(2^{k_1}x_1-j_1)\psi^{e_2}(2^{k_2}x_2-j_2). \tag{8.7}$$

Each of these functions has $L_2(\mathbb{R}^2)$ norm one. We let $\mathcal{L}$ denote the library of all complete orthonormal systems which can be made up from the functions in (8.7). In particular $\mathcal{L}$ will include the usual wavelet bases given in (7.19) and the hyperbolic basis (7.21), which is the tensor product of the univariate wavelet basis.

As a special case of the above library consider $\varphi=\chi_{[0,1)}$ and $\psi=H$, with H the Haar function. We approximate functions defined on the unit square $\Omega:=[0,1)^2$. The library $\mathcal{L}$ includes bases of the following type. We can take an arbitrary partition $\mathcal{P}$ of Ω into dyadic rectangles R. On each R we can take a standard or hyperbolic wavelet Haar basis. This library of bases is also closely related to the CART algorithm studied by Donoho (1997).

8.3. Approximation using n-terms from a dictionary

Suppose that $\mathbb{D}$ is a dictionary of functions from $\mathcal{H}$. It will be convenient to assume (without loss of generality in n-term approximation) that each $g\in\mathbb{D}$ has norm one ($\|g\|_{\mathcal{H}}=1$) and that $-g\in\mathbb{D}$ whenever $g\in\mathbb{D}$. One particular example of a dictionary is to start with an orthonormal basis B for $\mathcal{H}$ and to take $\mathbb{D}:=\{\pm b : b\in\mathcal{H}\}$. We shall say that this is the *dictionary generated by* B. For each $n\in\mathbb{N}$, we let $\Sigma_n:=\Sigma_n(\mathbb{D})$ denote the collection of all functions in H which can be expressed as a linear combination of at most n elements of $\mathbb{D}$. Thus each function $S\in\Sigma_n$ can be written in the form

$$S=\sum_{g\in\Lambda}c_g g,\quad \Lambda\subset\mathbb{D},\quad \#\Lambda\le n, \tag{8.8}$$

with the $c_g\in\mathbb{R}$. It may be possible to write an element from $\Sigma_n(\mathbb{D})$ in the form (8.8) in more than one way.

For a function $f\in H$, we define its approximation error

$$\sigma_n(f):=\sigma_n(f,\mathbb{D})_{\mathcal{H}}:=\inf_{S\in\Sigma_n}\|f-S\|_{\mathcal{H}}. \tag{8.9}$$

We shall be interested in estimates for σ_n (from above and below). For this purpose, we introduce the following way of measuring smoothness with respect to the dictionary $\mathbb{D}$.

For a general dictionary $\mathbb{D}$, and for any $\tau>0$, we define the class of functions

$$\mathcal{K}^o_\tau(\mathbb{D},M):=\left\{f=\sum_{g\in\Lambda}c_g g:\Lambda\subset\mathbb{D},\ \#\Lambda<\infty\quad\text{and}\quad\sum_{g\in\Lambda}|c_g|^\tau\le M^\tau\right\},$$

and we define $K_\tau(\mathbb{D}, M)$ as the closure (in $\mathcal{H}$) of $\mathcal{K}_\tau^o(\mathbb{D}, M)$. Furthermore, we define $\mathcal{K}_\tau(\mathbb{D})$ as the union of the classes $\mathcal{K}_\tau(\mathbb{D}, M)$ over all $M > 0$. For $f \in \mathcal{K}_\tau(\mathbb{D})$, we define the semi-norm

$$|f|_{\mathcal{K}_\tau(\mathbb{D})} \tag{8.10}$$

as the infimum of all M such that $f \in \mathcal{K}_\tau(\mathbb{D}, M)$. Notice that, when $\tau = 1$, $\mathcal{K}_1$ is the class of functions which are a convex combination of the functions in $\mathbb{D}$.

The case when $\mathbb{D}$ is generated by a basis B is instructive for the results that follow. In this case, n-term approximation from $\mathbb{D}$ is the same as n-term approximation from B which we have analysed in Sections 2 and 5. We have shown that if $1/\tau = \alpha + 1/2$, then f is in the approximation class $\mathcal{A}_\tau^\alpha(\mathbb{D})$ if and only if

$$\sum_k |\langle f, h_k \rangle|^\tau$$

is finite and this last expression is equivalent to $|f|^\tau_{\mathcal{A}_\tau(\mathcal{B})}$. In particular, this shows that

$$\sigma_n(f, \mathbb{D})_{\mathcal{H}} \le Cn^{-\alpha}|f|_{\mathcal{K}_\tau(\mathbb{D})} \tag{8.11}$$

in the special case that $\mathbb{D}$ is given by an orthonormal basis B.

We are now interested in understanding whether (8.11) holds for more general dictionaries $\mathbb{D}$. The results in the following section will show that (8.11) is valid for a general dictionary provided $\alpha \ge 1/2$. The first result of this type was due to Maurey (see Pisier (1980)) who showed that, in the case $\alpha = 1/2$, (8.11) is valid for any dictionary. An iterative algorithm to generate approximants from $\Sigma_n(\mathbb{D})$ that achieves this estimate (for $\alpha = 1/2$) was given by Jones (1992). For $\alpha > 1/2$, (8.11) is proved in DeVore and Temlyakov (1996). For $\alpha < 1/2$ ($1 \le \tau \le 2$) there seems to be no obvious analogue of (8.11) for general dictionaries.

8.4. Greedy algorithms

The estimate (8.11) can be proved for a general dictionary by using *greedy algorithms* (also known as *adaptive pursuit*). These algorithms are often used computationally as well. We shall mention three examples of greedy algorithms and analyse their approximation properties. In what follows, $\|\cdot\|$ is the norm in $\mathcal{H}$ and $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$.

The first algorithm, known as the *pure greedy algorithm*, can be applied for any dictionary $\mathbb{D}$. Its advantage is its simplicity. It begins with a target function f and successively generates approximants $G_m(f) \in \Sigma_m(\mathbb{D})$, $m = 1, 2, \ldots$. In the case that $\mathbb{D}$ is generated by an orthonormal basis B, $G_m(f)$ is a best m-term approximation to f.

If $f \in H$, we let $g = g(f) \in \mathbb{D}$ denote an element from $\mathbb{D}$ which maximizes $\langle f, g\rangle$:

$$\langle f, g(f)\rangle = \sup_{g\in D}\langle f, g\rangle. \tag{8.12}$$

We shall assume for simplicity that such a maximizer exists; if not, suitable modifications are necessary in the algorithms that follow. We define

$$G(f) := G(f, \mathbb{D}) := \langle f, g(f)\rangle g(f)$$

and

$$R(f) := R(f, D) := f - G(f).$$

Then, $G(f)$ is a best one-term approximation to f from $\mathbb{D}$ and $R(f)$ is the residual of this approximation.

Pure greedy algorithm. Initially, we set $R_0(f) := R_0(f, D) := f$ and $G_0(f) := 0$. Then, for each $m \geq 1$, we inductively define

$$\begin{aligned} G_m(f) &:= G_m(f, \mathbb{D}) := G_{m-1}(f) + G(R_{m-1}(f)), \\ R_m(f) &:= R_m(f, \mathbb{D}) := f - G_m(f) = R(R_{m-1}(f)). \end{aligned}$$

The pure greedy algorithm converges to f for each $f \in \mathcal{H}$ (see Davis, Mallat and Avellaneda (1997)). This algorithm is greedy in the sense that at each iteration it approximates the residual $R_m(f)$ as best possible by a single function from $\mathbb{D}$. If $\mathbb{D}$ is generated by an orthonormal basis, then it is easy to see that $G_m(f)$ is a best m-term approximation to f from $\mathbb{D}$ and

$$\sigma_m(f, \mathcal{B})_{\mathcal{H}} = \|f - G_m(f)\|_{\mathcal{H}} = \|R_m(f)\|_{\mathcal{H}}.$$

However, for general dictionaries, this is not the case, and in fact the approximation properties of this algorithm are somewhat in doubt, as we shall now describe.

For a general dictionary $\mathbb{D}$, the best estimate (proved in DeVore and Temlyakov (1996)) known for the pure greedy algorithm is that for each $f \in K_1(\mathbb{D})$ we have

$$\|f - G_m(f)\|_{\mathcal{H}} \leq |f|_{K_1(\mathbb{D})} m^{-1/6}. \tag{8.13}$$

Moreover, the same authors have given an example of a dictionary $\mathbb{D}$ and a function f which is a linear combination of two elements of $\mathbb{D}$ such that

$$\|f - G_m(f)\|_{\mathcal{H}} \geq C m^{-1/2}, \tag{8.14}$$

with C an absolute constant. In other words, for the simplest of functions f (which are in all of the smoothness classes $K_\tau(\mathbb{D})$), the pure greedy algorithm provides approximation of at most order $O(m^{-1/2})$. Thus, this algorithm cannot provide estimates like (8.11) for $\alpha > 1/2$.

There are modifications of the pure greedy algorithm with more favourable approximation properties. We mention two of these: the *relaxed greedy algorithm* and the *orthogonal greedy algorithm*.

Relaxed greedy algorithm. We define $R_0^r(f) := R_0^r(f, \mathbb{D}) := f$ and $G_0^r(f) := G_0^r(f, D) := 0$. For $m = 1$, we define $G_1^r(f) := G_1^r(f, \mathbb{D}) := G_1(f)$ and $R_1^r(f) := R_1^r(f, \mathbb{D}) := R_1(f)$. As before, for a function $h \in H$, let $g = g(h)$ denote a function from $\mathbb{D}$ which maximizes $\langle h, g \rangle$. Then, for each $m \geq 2$, we inductively define

$$\begin{aligned} G_m^r(f) &:= G_m^r(f, \mathbb{D}) := \left(1 - \frac{1}{m}\right) G_{m-1}^r(f) + \frac{1}{m} g(R_{m-1}^r(f)), \\ R_m^r(f) &:= R_m^r(f, \mathbb{D}) := f - G_m^r(f). \end{aligned}$$

Thus, the relaxed greedy algorithm is less greedy than the pure greedy algorithm. It makes only modest use of the greedy approximation to the residual at each step. The number $1/m$ appearing at each step is the relaxation parameter.

Algorithms of this type appear in Jones (1992), who showed that the relaxed greedy algorithm provides the approximation order

$$\|f - G_m^r(f)\| \leq C m^{-1/2}, \quad m = 1, 2, \ldots. \tag{8.15}$$

for any $f \in \mathcal{K}_1(\mathbb{D})$. Unfortunately, this estimate requires the knowledge that $f \in K_1(\mathbb{D})$. In the event that this information is not available – as would be the case in most numerical considerations – the choice of relaxation parameter $1/m$ is not appropriate.

The relaxed greedy algorithm gives a constructive proof that (8.11) holds for a general dictionary $\mathbb{D}$ in the case $\alpha = 1/2$. We shall discuss how to prove (8.11) in the next section. But first we want to put out on the table another variant of the greedy algorithm, called the orthogonal greedy algorithm, which removes some of the objections to the choice of the relaxation parameter in the relaxed greedy algorithm.

To motivate the orthogonal greedy algorithm, let us return for a moment to the pure greedy algorithm. This algorithm chooses functions $g_j := G(R_j(f))$, $j = 1, \ldots, m$, to use in approximating f. One of the deficiencies of the algorithm is that it does not provide the best approximation from the span of $g_1, \ldots, g_m$. We can remove this deficiency as follows.

If H_0 is a finite-dimensional subspace of H, we let P_{H_0} be the orthogonal projector from H onto H_0, that is, $P_{H_0}(f)$ is the best approximation to f from H_0.

Orthogonal greedy algorithm. We define $R_0^o(f) := R_0^o(f, D) := f$ and $G_0^o(f) := G_0^o(f, D) := 0$. Then, for each $m \geq 1$, we inductively define

$$H_m := H_m(f) := \operatorname{span}\{g(R_0^o(f)), \ldots, g(R_{m-1}^o(f))\},$$

$$\begin{aligned} G_m^o(f) &:= G_m^o(f,D) := P_{H_m}(f), \\ R_m^o(f) &:= R_m^o(f,D) := f - G_m^o(f). \end{aligned}$$

Thus, the distinction between the orthogonal greedy algorithm and the pure greedy algorithm is that the former takes the best approximation by linear combinations of the functions $G(R_0(f)), \ldots, G(R_{m-1}(f))$ available at each iteration. The first step of the orthogonal greedy algorithm is the same as the pure greedy algorithm. However, they will generally be different at later steps.

DeVore and Temlyakov (1996) have shown (as will be discussed in more detail in the next section) that the orthogonal greedy algorithm satisfies the estimate

$$\|f - G_m^o(f,D)\|_{\mathcal{H}} \le |f|_{K_1(\mathbb{D})} m^{-1/2}. \tag{8.16}$$

Thus, the orthogonal greedy algorithm gives another constructive proof that (8.11) holds for a general dictionary $\mathbb{D}$. However, one should note that the orthogonal greedy algorithm is computationally more expensive in the computation of the best approximation from H_m.

From (8.16), it is easy to prove the following theorem from DeVore and Temlyakov (1996).

Theorem 7 Let $\mathbb{D}$ be any dictionary, let $\alpha \ge 1/2$ and $1/\tau = \alpha + 1/2$. If $f \in K_\tau(\mathbb{D})$, then

$$\sigma_m(f,\mathbb{D})_{\mathcal{H}} \le C|f|_{K_\tau(D)} m^{-\alpha}, \quad m = 1, 2, \ldots, \tag{8.17}$$

where C depends on τ if τ is small.

We sketch the simple proof. It is enough to prove (8.17) for functions f which are a finite sum $f = \sum_j c_j g_j$, $g_j \in D$, with $\sum_j |c_j|^\tau \le M^\tau$. Without loss of generality we can assume that the c_j are positive and nonincreasing. We let $s_1 := \sum_{j=1}^n c_j g_j$ and $R_1 := f - s_1 = \sum_{j>n} c_j g_j$. Now,

$$c_n^\tau \le \frac{1}{n} \sum_{j=1}^n |c_j|^\tau \le \frac{M^\tau}{n}, \quad n = 1, 2, \ldots.$$

Hence, $c_j \le M n^{-1/\tau}$, $j > n$ and it follows that

$$\sum_{j>n} c_j = \sum_{j>n} c_j^{1-\tau} c_j^\tau \le M^{1-\tau} n^{1-1/\tau} \sum_{j>n} c_j^\tau \le M n^{1-1/\tau}.$$

This gives that R_1 is in $K_1(\mathbb{D}, Mn^{1-1/\tau})$. Using (8.16), there is a function s_2 which is a linear combination of at most n of the $g \in \mathbb{D}$ such that

$$\|f - (s_1 + s_2)\| = \|R_1 - s_2\| \le 2Mn^{1-1/\tau} n^{-1/2} = 2Mn^{-\alpha},$$

and (8.17) follows.

8.5. Further analysis of greedy algorithms

To determine the performance of a greedy algorithm, we try to estimate the decrease in error provided by one step of the pure greedy algorithm. Let $\mathbb{D}$ be an arbitrary dictionary. If $f \in \mathcal{H}$ and

$$\rho(f) := \langle f, g(f)\rangle / \|f\|_{\mathcal{H}}, \tag{8.18}$$

where as before $g(f) \in \mathbb{D}$ satisfies

$$\langle f, g(f)\rangle = \sup_{g \in D} \langle f, g\rangle,$$

then

$$R(f)^2 = \|f - G(f)\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{H}}^2 (1 - \rho(f)^2). \tag{8.19}$$

The larger $\rho(f)$ is, the better the decrease of the error in the pure greedy algorithm.

The following theorem from DeVore and Temlyakov (1996) estimates the error in approximation by the orthogonal greedy algorithm.

Theorem 8 Let $\mathbb{D}$ be an arbitrary dictionary in $\mathcal{H}$. Then, for each $f \in \mathcal{K}_1(\mathbb{D}, M)$ we have

$$\|f - G_m^o(f, \mathbb{D})\|_{\mathcal{H}} \le M m^{-1/2}. \tag{8.20}$$

Proof. We can assume that $M = 1$ and that f is in $\mathcal{K}_1^o(\mathbb{D}, 1)$. We let $f_m^o := R_m^o(f)$ be the residual at step m of the orthogonal greedy algorithm. Then, from the definition of this algorithm, we have

$$\|f_{m+1}^o\|_{\mathcal{H}} \le \|f_m^o - G(f_m^o, \mathbb{D})\|_{\mathcal{H}}.$$

Using (8.19), we obtain

$$\|f_{m+1}^o\|_{\mathcal{H}}^2 \le \|f_m^o\|_{\mathcal{H}}^2 (1 - \rho(f_m^o)^2). \tag{8.21}$$

Since $f \in \mathcal{K}_1^o(\mathbb{D}, 1)$, we can write $f = \sum_{k=1}^N c_k g_k$ with $c_k > 0$, $k = 1, \ldots, N$, and $\sum_{k=1}^N c_k = 1$. By the definition of the orthogonal greedy algorithm, $G_m^o(f) = P_{H_m} f$, and hence $f_m^o = f - G_m^o(f)$ is orthogonal to $G_m^o(f)$. Using this, we obtain

$$\|f_m^o\|_{\mathcal{H}}^2 = \langle f_m^o, f\rangle = \sum_{k=1}^N c_k \langle f_m^o, g_k\rangle \le \rho(f_m^o)\|f_m^o\|_{\mathcal{H}}.$$

Hence,

$$\rho(f_m^o) \ge \|f_m^o\|_{\mathcal{H}}.$$

Using this inequality in (8.21), we find

$$\|f_{m+1}^o\|_{\mathcal{H}}^2 \le \|f_m^o\|_{\mathcal{H}}^2 (1 - \|f_m^o\|_{\mathcal{H}}^2).$$

It is now easy to derive from this that $\|f_m^o\|_{\mathcal{H}}^2 \le 1/m$. $\square$

9. Lower estimates for approximation: n-widths

In this section, we shall try to understand better the limitations of linear and nonlinear approximation. Our analysis thus far has relied on the concept of approximation spaces. For example, we started with a sequence of linear or nonlinear spaces X_n and defined the approximation classes A_∞^α consisting of all functions that can be approximated with accuracy $O(n^{-\alpha})$ by the elements of X_n. We have stressed the importance of characterizing these approximation spaces in terms of something more classical such as smoothness spaces and in fact we have accomplished this in many settings. In this way, we have seen among other things that the classical nonlinear methods of approximation (like free knot splines or n-term approximation) outperform their counterparts in linear approximation.

To make these points more clearly, let us recall perhaps the simplest setting for the results we have presented. Namely, we consider $L_2(\Omega)$-approximation, $\Omega := [0,1)$, using the Haar wavelet H. Every function in $L_2(\Omega)$ has a decomposition

$$f = a\chi_{[0,1)} + \sum_{I \in D([0,1))} c_I(f) H_I, \quad c_I(f) := \langle f, H_I \rangle, \tag{9.1}$$

with the H_I normalized in $L_2(\Omega)$ and a the average of f over Ω.

In linear approximation, we take as our approximation to f the partial sum of the series (9.1) consisting of the first n terms with respect to the natural order of dyadic intervals (this is the ordering which gives priority first to size and then to orientation from left to right). For this approximation, we have seen that f is approximated in the norm of $L_2(\Omega)$ with accuracy $O(n^{-\alpha})$, $0 < \alpha < 1/2$, if and only if $f \in \mathrm{Lip}(\alpha, L_2(\Omega))$. The upper limit of $1/2$ for the characterization comes about because the Haar wavelet H is in $\mathrm{Lip}(1/2, L_2(\Omega))$ but in no higher-order Lipschitz space.

In nonlinear approximation, we approximated f by taking the partial sum of (9.1) which consists of the n terms with largest coefficients. It is clear that this form of approximation is at least as efficient as the linear approximation. We have seen that we can characterize the functions approximable with order $O(n^{-\alpha})$ by conditions on the wavelet coefficients that roughly correspond to smoothness of order α in L_τ with $1/\tau = \alpha + 1/2$ (see Remark 7.7 on page 118). In fact, it is easy to see that each function in $\mathrm{Lip}(\alpha, L_\gamma(\Omega))$ with $\gamma > \tau$ is approximated with this order by the nonlinear method.

Is this really convincing proof that nonlinear methods outperform linear methods? Certainly it shows that this nonlinear wavelet method outperforms the linear wavelet method. However, what can prevent some other linear method (not the wavelet method just described) from also containing the $\mathrm{Lip}(\alpha, L_\gamma(\Omega))$ classes in its $\mathcal{A}_\infty^\alpha$? There is a way of deciding whether this is possible by using the concept of n-widths, which we now describe.

There are many definitions of n-widths. For our purpose of measuring the performance of linear methods, the following definition of Kolmogorov is most appropriate. If X is a Banach space and K is a compact subset of X, we define

$$d_n(K) := \inf_{\dim(X_n)=n} \sup_{f\in K} E(f, X_n)_X, \tag{9.2}$$

where the infimum is taken over all n-dimensional linear spaces and of course $E(f, X_n)_X$ is the error in approximating f by the elements of X_n in the norm of X. So d_n measures the performance of the best n-dimensional space on the class K.

To answer our question posed above, we would like to know the n-width of the unit ball U_γ^α of $\mathrm{Lip}(\alpha, L_\gamma(\Omega))$ in $L_2(\Omega)$ (this unit ball is a compact subset of $L_2(\Omega)$ provided $\gamma > \tau = (\alpha + 1/2)^{-1}$). The Kolmogorov n-widths of Besov and Lipschitz balls are known and can be found, for example, in Chapter 14 of Lorentz, von Golitschek and Makovoz (1996). We shall limit our discussion to the results relevant to our comparison of linear and nonlinear approximation.

We fix the space $L_p(\Omega)$, $\Omega = [0, 1)$, where approximation is to take place. While we shall discuss only univariate approximation in this section, all results on n-widths hold equally well in the multivariate case. In Figure 8, we use our usual interpretation of smoothness spaces as points in the upper right quadrant to give information about the n-widths of the unit balls $U_r^\alpha(L_q(\Omega))$ of the Besov spaces $B_r^\alpha(L_q(\Omega))$. The shaded region of that figure corresponds to those Besov spaces whose unit ball has n-width $O(n^{-\alpha})$.

Several remarks will complete our understanding of Figure 8 and what it tells us regarding linear and nonlinear methods.

Remark 9.1 The n-width of $U_r^\alpha(L_q(\Omega))$ is never better than $\mathcal{O}(n^{-\alpha})$. In other words, once we know the smoothness index α of the space, this provides a limit as to how effective linear methods can be.

Remark 9.2 The sets $U_r^\alpha(L_p(\Omega))$ which correspond to the Besov spaces on the linear line (L) always have Kolmogorov n-width $\asymp n^{-\alpha}$. Thus, for these spaces the classical methods of approximation such as polynomials or fixed knot splines provide the best order of approximation for these classes.

Remark 9.3 For approximation in $L_p(\Omega)$, with $2 < p \le \infty$, and for $\alpha > 1/p$ there is always a certain range of q (depicted in Figure 8 by the shaded region) where the Kolmogorov n-width of $U_r^\alpha(L_q\Omega))$ is still $\asymp n^{-\alpha}$. This is a rather surprising result of Kashin (1977). We know that classical methods cannot provide this order of approximation because we have characterized their approximation classes $\mathcal{A}_\infty^\alpha(L_p(\Omega))$ and these classes do not contain general functions from $U_r^\alpha(L_q(\Omega))$ once $q < p$. So there are linear spaces with super approximation properties (which to a limited extent

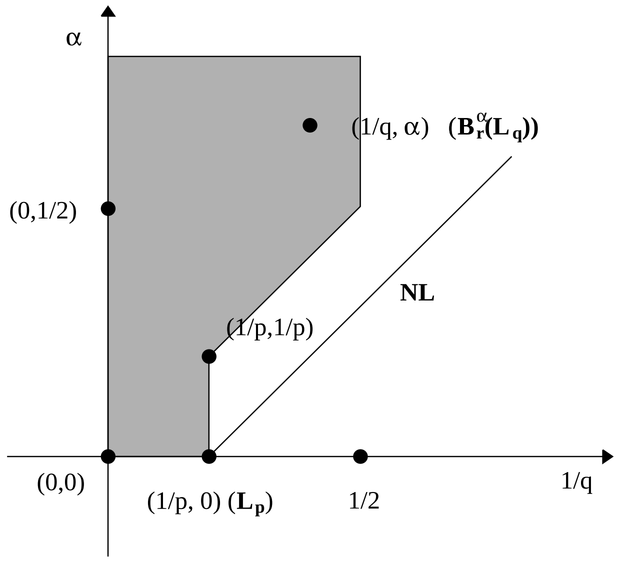

Fig. 8. Shaded region gives $(1/q, \alpha)$ such that $U_r^\alpha(L_q([0,1))$ has n-width of order $O(n^{-\alpha})$ in L_p, $2 \le p \le \infty$

mimic the advantages of nonlinear approximation). What are these spaces? Unfortunately these spaces are not known constructively. They are usually described by probabilistic methods. So, while their existence is known, we cannot put our hands on them and definitely can't use them numerically.

Remark 9.4 The range of q where the super linear spaces come into play always falls well short of the nonlinear line. Thus nonlinear methods always perform better than linear methods, in the sense that their approximation classes are strictly larger.

Remark 9.5 We have not depicted the case $p \le 2$ since in this case there are no Besov balls $U_r^\alpha(L_q(\Omega))$ which have the order $O(n^{-\alpha})$ save for the case $q \ge p$ which we already know from the classical linear theory.

Remark 9.6 Now, here is an important point that is sometimes misunderstood. It is not always safe to say that, for a specific target function, nonlinear methods will perform better than linear methods. Let us forget for a moment the super linear theory since it is not relevant in numerical situations anyway. Given f, there will be a maximal value of α – let's call it α_L – for which f is in $B_\infty^\alpha(L_p(\Omega))$. Then, we know that approximation from classical n-dimensional linear spaces will achieve an approximation rate $O(n^{-\alpha_L})$, but they can do no better. Let us similarly define α_N as the largest value of α for which f is in the space $B_\infty^{\alpha_N}(L_\gamma)$ for some $\gamma > (\alpha + 1/p)^{-1}$; then nonlinear methods such as n-term wavelet approximation will provide an approximation error $O(n^{-\alpha_N})$. If $\alpha_N > \alpha_L$, then certainly nonlinear methods outperform linear methods (at least asymptotically as $n \to \infty$). However, if $\alpha_L = \alpha_N$ then there is no gain in using nonlinear methods to approximate the target function f.

The questions we have posed for linear approximation can likewise be posed for nonlinear methods. For example, consider univariate approximation in $L_p(\Omega)$, $\Omega = [0,1)$. We know that classical nonlinear methods approximate functions in $B^\alpha_\infty(L_\gamma)$, $\gamma > (\alpha + 1/p)^{-1}$ with accuracy $n^{-\alpha}$. But can it be that other nonlinear methods do better? Questions of this type can be answered by introducing *nonlinear* n-widths.

There are several definitions of nonlinear n-widths, the most prominent of which is the Alexandrov width. However, we shall only be concerned with the manifold n-width, which was introduced by DeVore, Howard and Micchelli (1989), since it fits best with numerical methods. Let X be the space in which we shall measure error (we shall assume that X is equipped with a norm $\|\cdot\|_X$). By a (nonlinear) manifold $\mathcal{M}_n$ of dimension n, we shall mean the image of a continuous mapping $M : \mathbb{R}^n \to X$. (Thus our manifolds are *not* the manifolds of differential topology.) We shall approximate using the elements of $\mathcal{M}_n$. For each compact set $K \subset X$, we define the manifold width

$$\delta_n(K, X) := \inf_{M,a} \sup_{f \in K} \|f - M(a(f))\|_X, \tag{9.3}$$

where the infimum is taken over all manifolds of dimension n and all continuous parameter mappings $a : K \to \mathbb{R}^n$.

We make a couple of remarks which may help explain the nature of the width δ_n.

Remark 9.7 For any compact set, we can select a countable number of points which are dense in K and construct a one-dimensional manifold (a continuous piecewise linear function of $t \in \mathbb{R}$) passing through each of these points. Thus, without the restriction that the approximation arises through a continuous parameter selection a, we would always have $\delta_n(K) = 0$.

Remark 9.8 The function a also guarantees stability of the approximation process. If we perturb f slightly the continuity of a guarantees that the parameters $a(f)$ only change slightly.

The nonlinear widths of each of the Besov balls $U^\alpha_r(L_\tau(\Omega))$ in the space $L_p(\Omega)$ are known. If this ball is a compact subset of $L_p(\Omega)$, then the nonlinear n-width is

$$\delta_n(U^\alpha_r(L_\tau(\Omega)) \asymp n^{-\alpha}, \quad n \to \infty. \tag{9.4}$$

This shows, therefore, that we cannot obtain a better approximation order for these balls than what we obtain via n-term wavelet approximation. However, n-term approximation, as it now stands, is not described as one of the procedures appearing in (9.3). However, this requires only a little massaging. Using certain results from topology, DeVore, Kyriazis, Leviatan and Tikhomirov (1993) have shown nonlinear approximation in terms of soft thresholding of the coefficients can be used to describe an approximation process which provides the upper estimate in (9.3). We shall not go into the details of this construction.

On the basis of the evidence we have thus far provided about linear and nonlinear methods, is it safe to conclude that the nonlinear methods such as n-term wavelet approximation are superior to other nonlinear methods? The answer is definitely not. We only know that if we classify functions according to their Besov smoothness, then for this classification no other nonlinear methods can do better. On the other hand, each nonlinear method will have its approximation classes and these need not be Besov spaces. A case in point where we have seen this is the case of approximation in a Hilbert space by n terms of an orthonormal basis. In this setting, we have seen that the approximation classes depend on the basis and that smoothness of a function for this type of approximation should be viewed as decay of the coefficients with respect to the basis. This will generally not be a Besov space. In other words, there are other ways to measure smoothness in which wavelet performance will not be optimal.

Our discussion thus far has not included lower estimates for optimal basis selection or n-term approximation from a dictionary. We do not know of a concept of widths that properly measures the performance of these highly nonlinear methods of approximation. This is an important open problem in nonlinear approximation because it would shed light on the role of such methods in applications such as image compression (see the section below).

Finally, we want to mention the *VC* dimension of Vapnik and Chervonenkis (see the book of Vapnik (1982)). The VC dimension measures the size of nonlinear sets of functions by looking at the maximum number of sign alternations of its elements. It has an important role in statistical estimation but has not been fully considered in approximation settings. The paper of Mairov and Ratasby (1998) uses VC dimension to define a new n-width and analyses the widths of Besov balls. Their results are similar to those above for nonlinear widths.

10. Applications of nonlinear approximation

Nonlinear methods have found many applications both numerical and analytical. The most prominent of these have been to image processing, statistical estimation, and the numerical and analytic treatment of differential equations. There are several excellent accounts of these matters: see Mallat (1998) for image processing; Donoho and Johnstone (1994), Donoho, Johnstone, Kerkyacharian and Picard (1996) for statistical estimation; Dahmen

(1997) and Dahlke, Dahmen and DeVore (1997) for applications to PDEs. We shall limit ourselves to a broad outline of the use of nonlinear approximation in image processing and PDEs.

10.1. Image processing

We shall discuss the processing of digitized grey-scale images. Signals, colour images, and other variants can be treated similarly but have their own peculiarities. A digitized grey-scale image $\mathcal{I}$ is an array of numbers (called *pixel values*) which represent the grey scale. We assume 8-bit grey-scale images, which means the pixel values range from 0 (black) to 255 (white). We shall also assume (only for the sake of specificity) that the array consists of 1024×1024 pixel values. Given such images, the generic problems of image processing are: compression, noise reduction, feature extraction, and object recognition.

To utilize techniques from mathematical analysis in image processing, it is useful to have a model for images as functions. One such model is to assume that the pixel values are obtained from an underlying intensity function f by averaging over dyadic squares. In our case, the dyadic squares are those in $D_m := D_m(\Omega)$, $\Omega := [0,1)^2$, with $m = 10$, thus resulting in 1024 squares and the same number of pixel values. We denote the pixel values by

$$p_I = \tfrac{1}{|I|} \int_I f(x)\,\mathrm{d}x, \quad I \in D_m. \tag{10.1}$$

Of course, there is more than one function f with these pixel values. Since the pixel values are integers, another possibility would be to view them as integer quantizations of the averages of f. In other words, other natural models may be proposed. But the main point is to visualize the image as obtained from an intensity function f.

Compression

A grey-scale image $\mathcal{I}$ of the type described is represented by its pixel array, $\mathcal{I} \sim (p_I)_{I \in D_m}$, which is a file of size one megabyte. For the purposes of transmission, storage, or other processing, we would like to represent this image with fewer bits. This can be accomplished in two ways. *Lossless encoding* of the image uses techniques from information theory to encode the image in fewer bits. The encoded image is identical to the original; in other words the process of encoding is reversible. *Lossy compression* replaces the original image by an approximation. This allows for more compression but with the potential loss of fidelity. Lossless encoders will typically result in compression factors of 2–1 which means the original file is reduced by half. Much higher compression factors can be obtained in lossy compression with no perceived degradation of the original image (images compressed by factors of 10–1 are typically indistinguishable from the original).

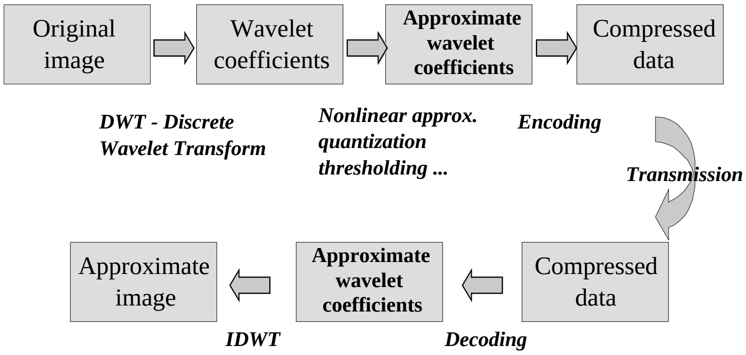

Fig. 9. Schematic of a typical wavelet-based compression algorithm

We can use the techniques of approximation theory and functional analysis for lossy compression. We view the intensity function as our target function and consider methods for approximating it from the pixel values. Wavelet-based methods proceed as follows.

We choose a multivariate scaling function φ and represent the image by the series

$$\mathcal{I} \sim \sum_{I \in D_m} p_I \varphi_I. \tag{10.2}$$

Here p_I, $I \in D_m$, are some appropriate extension of the pixel values. (When using wavelets other than Haar, one has to do some massaging near the boundary, which we shall not discuss.) We use the Fast Wavelet Transform to convert pixel values to wavelet coefficients. This gives the wavelet representation of $\mathcal{I}$:

$$\mathcal{I} \sim P_0 + \sum_{k=0}^{m-1} \sum_{I \in D_k} \sum_{e \in E} a_I^e \psi_I^e, \tag{10.3}$$

where P_0 consists of all the scaling function terms from level 0, and the other notation conforms to our multivariate wavelet notation of Section 7 (see (7.20)).

The problem of image compression is then viewed as nonlinear wavelet approximation and the results of Section 7 can be employed. Figure 9 gives a schematic of typical compression algorithms. We use thresholding to obtain a compressed file $(\tilde{a}_I^e)$ of wavelet coefficients which correspond to a compressed image $\tilde{I}$. The compressed coefficient file is further compressed using a lossless encoder. The encoded compressed file is our compressed representation of the original image. We can reverse the process. From the encoded compressed file of wavelet coefficients, we apply a decoder and then the Inverse Fast Wavelet Transform to obtain the pixel values of the compressed image $\tilde{I}$. The following remarks will help clarify the role of nonlinear approximation in this process.

Remark 10.1 We apply nonlinear wavelet approximation in the form of thresholding (Section 7.8). We choose a value of p (corresponding to the L_p space in which we are to measure error) and retain all coefficients that satisfy $\|a_I^e \psi_I^e\|_{L_p} > \epsilon$. We replace by zero all coefficients for which $\|a_I^e \psi_I^e\|_{L_p} \le \epsilon$. Soft thresholding can also be used in place of hard thresholding. This gives compressed wavelet coefficients $\bar{a}_I^e$. The larger we choose ϵ the more coefficients $\tilde{a}_I^e$ will be zero. In most applications, p is chosen to be 2. Larger values of p will emphasize edges, smaller values emphasize smoothness.

Remark 10.2 Further compression, in terms of number of bits, can be attained by quantizing the compressed wavelet coefficients. This means that $\bar{a}_I^e$ is replaced by a number $\tilde{a}_I^e$ which requires fewer bits in its binary representation. Quantization can be combined with thresholding by finding $\tilde{a}_I^e$ with the fewest bits which satisfies $\|(a_I^e - \tilde{a}_I^e)\psi_I^e\|_{L_p} \le \epsilon$.

Remark 10.3 The wavelet coefficient file consisting of the $\tilde{a}_I^e$ is further compressed by using a lossless encoder such as run length encoding or arithmetic encoding. The position of the coefficients must be encoded as well as their value. This can be done by keeping the entire array of coefficients in natural order (which will necessarily have many zero entries) or separately encoding positions.

Remark 10.4 The most efficient wavelet-based compression algorithms, such as the zero tree encoders (see Shapiro (1993) or Xiong, Ramchandran and Orchard (1997)) or bitstream encoder (see Gao and Sharpley (1997)), take advantage of the spatial correlation of the wavelet coefficients. For example, if we represent the coefficients by means of quadtrees with each node of the tree corresponding to one of the dyadic square I appearing in (10.3), then there will be many subtrees consisting only of zero entries, and one tries to encode these efficiently.

Remark 10.5 We can measure the efficiency of compression by the error

$$\sigma_n := \|I - \tilde{I}\|_{L_p}, \tag{10.4}$$

where n is the number of nonzero coefficients in the compressed wavelet file for $\tilde{I}$. Nonlinear approximation theory gives a direct relation between the rate of decrease of σ_n and the smoothness of the intensity function f. For example, consider approximation in L_2. If f is in the Besov class $B_\tau^\alpha(L_\tau)$, $1/\tau = \alpha/2 + 1/2$, then $\sigma_n \le Cn^{-\alpha/2}$. Indeed, assuming this smoothness for f, one can show that the function in (10.2) inherits this smoothness (see Chambolle, DeVore, Lee and Lucier (1998)) and therefore the claim follows from the results of Sections 7.6–7.7. An inverse theorem provides converse statements that deduce smoothness of the intensity function from the rate of compression. However, for these converse results one must think of varying m, that is, finer and finer pixel representations. The point is that one can associate to each image a smoothness index α which measures its smoothness in the above scale of Besov spaces, and relate this directly with efficiency of wavelet compression (DeVore, Jawerth and Lucier 1992).

Remark 10.6 In image compression, we are not interested in the number of nonzero coefficients of the compressed image *per se*, but rather the number of bits in the encoded coefficient file. This leads one to consider the error

$$\rho_n := \|I - \tilde{I}\|_{L_p}, \tag{10.5}$$

where n is the number of bits in the encoded file of wavelet coefficients for $\tilde{I}$. It has recently been shown by Cohen, Daubechies, Guleryuz and Orchard (1997) that a similar analysis to that developed here for nonlinear wavelet approximation exists for the error ρ_n. For example, they show that if a univariate intensity function f is in the Besov space $B^{\alpha}_{\infty}(L_q)$, with $q > \alpha + 1/2$, then with a proper choice of encoder one has $\rho_N \leq N^{-\alpha}$. This matches the error rate σ_n in terms of the number of coefficients. Related results hold in a stochastic setting (see Mallat and Falzon (1997) and Cohen, Daubechies, Guleryuz and Orchard (1997)).

Remark 10.7 Adaptive basis selection for the wavelet packet library has been used successfully in compression. Most applications have been to signal processing (in particular speech signals). There is, however, the interesting application of compressing the FBI fingerprint files. Rather than use a different basis for each file, the current algorithms choose one basis of the wavelet packet library chosen by its performance on a sample collection of fingerprint files.

Noise reduction

Noise reduction is quite similar to compression. If an image is corrupted by noise then the noisy pixel values will be converted to noisy wavelet coefficients. Large wavelet coefficients are thought to carry mostly signal and should be retained; small coefficients are thought to be mostly noise and should be thresholded to zero. Donoho and Johnstone have put forward algorithms for noise reduction (called wavelet shrinkage) which have elements similar to the above theory of compression. We give a brief description of certain aspects of this theory as it relates to nonlinear approximation. We refer the reader to Donoho, Johnstone, Kerkyacharian and Picard (1996), and the papers referenced therein, for a more complete description of the properties of wavelet shrinkage.

Wavelet-based noise reduction algorithms are applied even when the noise characteristics are unknown. However, the theory has its most complete description in the case that the pixel values are corrupted by Gaussian noise. This means we are given a noisy image $\tilde{\mathcal{I}} = \mathcal{I} + \mathcal{N}$ with noisy pixel values

$$\tilde{p}_I = p_I + \eta_I, \tag{10.6}$$

where the p_I are the original (noise-free) pixel values and the η_I are independent, identically distributed Gaussians with mean 0 and variance σ_0^2. If we choose an orthonormal wavelet basis for $L_2(\Omega)$, $\Omega = [0,1)^2$, then the wavelet coefficients computed from the $\tilde{p}_I$ will take the form

$$\tilde{c}_I^e = c_I^e + \epsilon_I^e, \tag{10.7}$$

where c_I^e are the original wavelet coefficients of $\mathcal{I}$ and ϵ_I^e are independent, identically distributed Gaussians with variance $\sigma_0^2 2^{-2m}$. Wavelet shrinkage with parameter $\lambda > 0$ replaces $\tilde{c}_I^e$ by the shrunk coefficients $s_\lambda(c_I^e)$ where

$$s_\lambda(t) := \begin{cases} (|t| - \lambda)\text{sign } t, & \lambda < t, \\ 0, & |t| \le \lambda, \end{cases} \tag{10.8}$$

Thus, large coefficients (*i.e.*, those larger than λ in absolute value) are shrunk by an amount λ and small coefficients are shrunk to zero. We denote the function with these wavelet coefficients by

$$f_\lambda := P_0 + \sum_{j=0}^{m-1} \sum_{I \in D_j} \sum_{e \in E} s_\lambda(\tilde{c}_I^e)\psi_{I,e}, \tag{10.9}$$

with the term P_0 incorporating the scaling functions from the coarsest level. We seek a value of λ which minimizes the expected error

$$E(\|f - f_\lambda\|^2_{L_2(\Omega)}). \tag{10.10}$$

Donoho and Johnstone propose the parameter choice $\lambda^* = \sqrt{2 \ln 2^m} 2^m \sigma_0$ and show its near optimality in several statistical senses. One of the extremal problems studied by them, as well as by DeVore and Lucier (1992), is the following. We assume that the original image intensity function f comes from the the Besov space $B_\tau^\alpha(L_\tau(\Omega))$, with $\tau = (\alpha/2 + 1/2)^{-1}$. We know that these spaces characterize the approximation space $A_\tau^\alpha(L_2(\Omega))$ for bivariate nonlinear wavelet approximation. It can be shown that the above choice of λ gives the noise reduction

$$E(\|f - f_\lambda\|^2) \le C(\lambda) \|f\|^\tau_{B_\tau^\alpha(L_\tau(\Omega))} [\sigma_0 2^{-m}]^{2-\tau}. \tag{10.11}$$

The choice of $\lambda = \lambda^*$ gives an absolute constant $c(\lambda^*)$. A finer analysis of this error was given by Chambolle, DeVore, Lee and Lucier (1998) and shows that choosing the shrinkage parameter to depend on α will result in an improved error estimate.

Significant improvements in noise reduction (at least in the visual quality of the images) can be obtained by using the technique of cycle spinning, as proposed by Coifman and Donoho (1995). The idea behind their method can be described by the following analysis of discontinuities of a univariate function g. The performance of wavelet-based compression and noise reduction algorithms depends on the position of the discontinuities. If a discontinuity of g occurs at a coarse dyadic rational, say $1/2$, it will affect only a few wavelet coefficients. These coefficients will be the ones that are changed by shrinking. On the other hand, if the discontinuity occurs at a fine level rational binary, say 2^{-m}, then all coefficients will feel this discontinuity and can potentially be affected by shrinkage. This less favourable situation can be circumvented by translating the image, so that the discontinuity appears at a coarse binary rational, and then applying wavelet shrinkage to the translated image. The image is shifted back to the original position to obtain the noise reduced image. Since it is not possible to anticipate the position of the discontinuities, Coifman and Donoho propose averaging over all possible shifts. The result is an algorithm that involves $O(m2^{2m})$ computations.

Feature extraction and object recognition
The time-frequency localization of wavelets allows for the extraction of features such as edges and texture. These can then be utilized for object recognition by matching the extraction to a corresponding template for the object to be extracted. Edges and other discontinuities are identifiable by the large wavelet coefficients. These occur at every dyadic level. Retention of high frequency (*i.e.*, the highest level) coefficients is like an artist's sketch of an image.

Feature extraction has been a prominent application of adaptive basis selection and approximation from a dictionary. A dictionary of waveforms is utilized which is robust enough to allow the feature to be approximated with a few terms. Examples are the Gabor functions mentioned in Section 8. In some cases, an understanding of the physics of wave propagation can allow the designing of dictionaries appropriate for the features to be extracted. A good example of this approach in the context of Synthetic Aperture Radar is given by McClure and Carin (1997). The use of adaptive basis selection for feature extraction is well represented in the book of Wickerhauser (1994). The application of greedy algorithms and approximation from dictionaries is discussed in detail in the book of Mallat (1998). Other techniques based on wavelet decompositions can be found in DeVore, Lucier and Yang (1996) (in digital mammography), and DeVore et al. (1997) (in image registration).

10.2. Analytical and numerical methods for PDEs

To a certain extent, one can view the problem of numerically recovering a solution u to a PDE (or system of PDEs) as a problem of approximating the target function u. However, there is a large distinction in the information available about u in numerical computation versus approximation theory. In approximation theory one views information such as point values of a function or wavelet coefficients as known, and constructs methods of approximation using this information. However, in numerical methods for PDEs, the target function is unknown except through the PDE. Thus, the information the approximation theorist wants and loves so much is not available except through numerical computation. In spite of this divergence of viewpoints, approximation theory can be very useful in numerical computation in suggesting numerical algorithms and, more importantly, to clarify the performance expected from linear and nonlinear numerical methods.

Adaptive methods are commonly used for numerically resolving PDEs. These methods can be viewed as a form of nonlinear approximation with the target function the unknown solution u to the PDE. Most adaptive numerical methods have not even been shown to converge and certainly have not been *theoretically proven* to have numerical efficiency over linear methods. Nevertheless, they have been very successful in practice and their effficiency has been experimentally established.

Nonlinear approximation can be very useful in understanding when and how adaptive numerical methods should be used. For example, from the analysis put forward in this paper, we know that adaptive piecewise polynomial methods, as well as the n-term wavelet approximation methods, have increased efficiency over linear methods when the target function u has certain types of singularities; specifically, singularities that would destroy the smoothness of u in the Sobolev scale but would not impair its smoothness in the Besov scale for nonlinear approximation.

To be more precise, suppose that u is to be approximated in the $L_p(\Omega)$ norm with Ω a domain in $\mathbb{R}^d$. Let α_L be the largest value of α such that u is in the Besov space $B^\alpha_\infty(L_p(\Omega))$. We know that u can be approximated by linear methods such as piecewise polynomial or linear wavelet approximation with accuracy $O(n^{-\alpha_L/d})$, with n the dimension of the linear space. However, we do not know (unless we prove it) whether our particular numerical method has this efficiency. If we wish to establish the efficiency of our particular linear numerical method, we should seek an estimate of the form

$$\|u - u_n\|_{L_p(\Omega)} \le C|u|_{B^{\alpha_L}_\infty(L_p(\Omega))} n^{-\alpha_L/d}, \tag{10.12}$$

where u_n is the approximate solution provided by our numerical method. In many papers, $W^{\alpha_L}(L_p(\Omega))$ is used in place of $B^{\alpha_L}_\infty(L_p(\Omega)$. The form of such estimates is familiar to the numerical analyst in finite element methods where such estimates are known in various settings (especially in the case $p = 2$ since this can be related to the energy norm).

Note that n is related to the numerical effort needed to compute the approximant. However, the number of computations needed to compute an approximant with this accuracy may exceed Cn. This may be the case, for example, in solving elliptic equations with finite element methods, since the coefficients of the unknown solution must be computed as a solution to a matrix equation of size $n \times n$.

We can do a similar analysis for nonlinear methods. According to the results reviewed in this article, the appropriate scale of Besov spaces to gauge the performance of nonlinear algorithms are the $B_q^\alpha(L_q(\Omega))$ where $1/q = \alpha/d + 1/p$ (see Figure 3 in the case $d = 1$). Let α_N be the largest value of α such that u is in the Besov space $B_q^\alpha(L_q(\Omega))$, $1/q = \alpha/d + 1/p$. If $\alpha_N > \alpha_L$, then nonlinear approximation will be more efficient than linear approximation in approximating u and therefore the use of nonlinear methods is completely justified. However, there still remains the question of how to construct a nonlinear algorithm that approximates u with the efficiency $O(n^{-\alpha_N/d})$. If we have a particular nonlinear numerical method in hand and wish to analyse its efficiency, then the correct form of an error estimate for such a nonlinear algorithm would be

$$\|u - u_n\|_{L_p(\Omega)} \leq C|u|_{B_q^{\alpha_N}(L_q(\Omega))} n^{-\alpha_N/d}, \quad 1/q = \alpha_N/d + 1/p. \tag{10.13}$$

How could we decide beforehand whether nonlinear methods offer a benefit over linear methods? This is the role of regularity theorems for PDEs. A typical regularity theorem infers the smoothness of the solution u to a PDE from information such as the coefficients, inhomogeneous term, initial conditions, or boundary conditions. We shall discuss this in a little more detail in a moment, but for now we want to make the point of what form these regularity theorems should take. The most common regularity theorems are in the form of Sobolev regularity and are compatible with the linear theory of numerical methods. Much less emphasis has been placed on the regularity in the nonlinear scale of Besov spaces but this is exactly what we need for an analysis of adaptive, or other nonlinear, algorithms.

To go a little further in our discussion, we shall consider two model problems, one hyperbolic and the other elliptic, to elucidate the points discussed above.

Conservation laws

Consider the scalar univariate conservation law

$$\begin{cases} u_t + f(u)_x = 0, & x \in \mathbb{R},\ t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{10.14}$$

where f is a given flux, u_0 a given initial condition, and u is the sought-after solution. This is a well studied nonlinear transport equation with transport velocity $a(u) = f'(u)$ (see, for instance, the book of Godlewski and Raviart (1991)). We shall assume that the flux is strictly convex, which means the transport velocity is strictly increasing. The important fact for us is that, even when the initial condition u_0 is smooth, the solution $u(\cdot,t)$ will develop spontaneous shock discontinuities at later times t.

The proper setting for the analysis of conservation laws is in L_1 and, in particular, the error of numerical methods should be measured in this space. Thus, concerning the performance of linear numerical methods, the question arises as to the possible values of the smoothness parameter α_L of $u(\cdot,t)$ as measured in L_1. It is known that, if the initial condition u_0 is in $BV = \mathrm{Lip}(1, L_1)$, then the solution u remains in this space for all later time $t > 0$. However, since this solution develops discontinuities, no matter how smooth the initial condition is, the Sobolev embedding theorem precludes u being in any Besov space $B^\alpha_\infty(L_1))$ for any $\alpha > 1$. This means that the largest value we can expect for α_L is $\alpha_L = 1$. Thus, the optimal performance we can expect from linear methods of approximation is $O(n^{-1})$, with n the dimension of the linear spaces used in the approximation. Typical numerical methods utilize spaces of piecewise polynomials on a uniform mesh with mesh length h and the above remarks mean that the maximum efficiency we can expect for numerical methods is $O(h)$, $h \to 0$. In reality, the best proven estimates are $O(\sqrt{h})$ under the assumption that $u_0 \in \mathrm{Lip}(1, L_1)$. This discrepancy between the possible performance of numerical algorithms and the actual performance is not unusual. The solution is known to have sufficient regularity to be approximated, for example, by piecewise constants with uniform mesh h to accuracy $O(h)$, but algorithms which capture this accuracy are unkown.

To understand the possible performance of nonlinear methods such as moving grid methods, we should estimate the smoothness of the solution in the nonlinear Besov scale $B^\alpha_\tau(L_\tau))$, $1/\tau = \alpha + 1$, corresponding to approximation in the L_1-norm. A rather surprising result of DeVore and Lucier (1990) shows that, starting with a smooth initial condition u_0, the solution u will be in each of these Besov spaces for all $\alpha > 0$. In other words, depending on the smoothness of u_0, α_N can be arbitrarily large. This means that nonlinear methods such as moving grid methods could provide arbitrarily high efficiency. In fact, such algorithms, based on piecewise polynomial approximation, can be constructed using the method of characteristics (see Lucier (1986) for the case of piecewise linear approximation).

Unfortunately, the situation concerning numerical methods for multivariate conservation laws is not as clear. While the linear theory goes through almost *verbatim*, the nonlinear theory is left wanting. The proper form of nonlinear approximation in the multivariate case would most likely be by piecewise polynomials on free triangulations. As we have noted earlier, it is an unsolved problem in nonlinear approximation to describe the smoothness conditions that govern the efficiency of this type of approximation. For a further discussion of the multivariate case see DeVore and Lucier (1996).

Because of their unique ability to detect singularities in a function, wavelet methods seem a natural candidate for numerical resolution of solutions to conservation laws. However, it is not yet completely clear how wavelets should be used in numerical solvers. Attempts to use wavelets directly in a time-stepping solver have not been completely effective. Ami Harten (1994) and his collaborators have suggested the use of wavelets to compress the computations in numerical algorithms. For example, he proposes the use of a standard time-stepping solver, such as Godunov, based on cell averages for computing the solution at time step t_{n+1} from the numerically computed solution at time step t_n, but to utilize wavelet compression to reduce the number of flux computations in the solution step.

Elliptic equations

An extensive accounting of the role of linear and nonlinear approximation in the solution of elliptic problems is given in Dahmen (1997) and Dahlke, Dahmen and DeVore (1997). We shall therefore limit ourselves to reiterating a couple of important points about the role of regularity theorems and the form of nonlinear estimates. We consider the model problem

$$\begin{aligned} \triangle u &= f \quad \text{on} \quad \Omega \subset \mathbb{R}^d, \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned} \tag{10.15}$$

of Poisson's equation on a domain $\Omega \subset \mathbb{R}^d$ with zero boundary conditions. We are interested in numerical methods for recovering the solution to (10.15) and, in particular, in the question of whether nonlinear methods such as adaptive solvers are of any benefit. We shall also limit our discussion to estimating error in the L_2-norm, although various results are known for general p.

Consider first the case where $f \in L_2(\Omega)$ and Ω has a smooth boundary. Then, the solution u to (10.15) has smoothness $W^2(L_2(\Omega))$. In our previous notation, this means that $\alpha_L = 2$. In general, the solution will not have higher smoothness in the nonlinear Besov scale $B_q^\alpha(L_q(\Omega))$, $1/q = \alpha/d + 1/2$, for L_2 approximation. Therefore $\alpha_N = 2$ and there is no apparent advantage to nonlinear methods. The solution can be approximated by linear spaces of piecewise polynomials of dimension n to accuracy $O(n^{-2/d})$. This accuracy can actually be achieved by finite element methods using uniformly refined partitions. There is no evidence to suggest any better performance using adaptive methods.

If the boundary $\partial\Omega$ of Ω is not smooth then the solutions (10.15) have singularities due to corners or other discontinuities of $\partial\Omega$ (see, for instance, Kondrat'ev and Oleinik (1983)). Regularity theory in the case of a non-smooth boundary is a prominent area of PDEs. For some of the deepest and most recent results see Jerison and Kenig (1995). For example, on a general Lipschitz domain, we can only expect that the solution u to (10.15) is in the Sobolev space $W^{3/2}(L_2(\Omega))$. Thus, in the notation given earlier in this section, we will only have $\alpha_L = 3/2$.

Because of the appearance of singularities due to the boundary, adaptive numerical techniques are suggested for numerically recovering the solution u. We understand that to justify the use of such methods we should determine the regularity of the solution in the scale of Besov spaces $B_q^\alpha(L_q(\Omega))$, $1/q = \alpha/d + 1/2$. Such regularity has been studied by Dahlke and DeVore (1997). They prove, for example, that, for $d = 2, 3, 4$, we have $u \in B_q^\alpha(L_q)$, $1/q = \alpha/d + 1/2$, for each $\alpha < 2$. In other words, $\alpha_N > \alpha_L$ and the use of adaptive methods is completely justified. There are also more general results which apply for any $d > 1$ and show that we always have $\alpha_N > \alpha_L$.

We reiterate that the above results on regularity of elliptic equations only indicate the possibility of constructing nonlinear methods with higher efficiency. It remains a difficult problem to construct adaptive methods and prove that they exhibit the increased accuracy indicated by the approximation theory. The aim is to construct numerical methods that provide the error estimate (10.13). We refer the reader to Dahmen (1997) for a comprehensive discussion of what is known about adaptive methods for elliptic equations.

Acknowledgements
The author thanks Professors de Boor, Cohen, Oskolkov, Petrushev and Temlyakov for their valuable suggestions concerning this survey.

REFERENCES

R. A. Adams (1975), *Sobolev Spaces*, Academic Press, New York.

I. Babuška and M. Suri (1994), 'The p and h versions of the finite element method: basic principles and properties', *SIAM Review* **36**, 578–632.

G. A. Baker, Jr. (1975), *Essentials of Padé Approximants*, Academic Press, New York.

C. Bennett and R. Sharpley (1988), *Interpolation of Operators*, Academic Press, New York.

J. Bergh and J. Löfström (1976), *Interpolation Spaces: An Introduction*, Springer, Berlin.

J. Bergh and J. Peetre (1974), 'On the spaces V_p $(0 < p \le \infty)$', *Boll. Unione Mat. Ital.* **10**, 632–648.

M. Birman and M. Solomyak (1967), 'Piecewise polynomial approximation of functions of the class W_p^α', *Mat. Sbornik* **2**, 295–317.

C. de Boor (1973), 'Good approximation by splines with variable knots', in *Spline Functions and Approximation* (A. Meir and A. Sharma, eds), Birkhäuser, Basel, pp. 57–72.

C. de Boor, R. DeVore and A. Ron (1993), 'Approximation from shift invariant spaces', *Trans. Amer. Math. Soc.* **341**, 787–806.

L. Brown and B. Lucier (1994), 'Best approximations in L_1 are best in L_p, $p < 1$, too', *Proc. Amer. Math. Soc.* **120**, 97–100.

Yu. Brudnyi (1974), 'Spline approximation of functions of bounded variation', *Soviet Math. Dokl.* **15**, 518–521.

A. P. Calderón (1964*a*), 'Intermediate spaces and interpolation: the complex method', *Studia Math.* **24**, 113–190.

A. P. Calderón (1964*b*), 'Spaces between L_1 and L_∞ and the theorem of Marcinkieiwicz: the complex method', *Studia Math.* **26**, 273–279.

A. Chambolle, R. DeVore, N.-Y. Lee and B. Lucier (1998), 'Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage', *IEEE Trans. Image Processing* **7**, 319–335.

A. Cohen, I. Daubechies and J.-C. Feauveau (1992), 'Biorthogonal bases of compactly supported wavelets', *Comm. Pure Appl. Math.* **43**, 485–560.

A. Cohen, I. Daubechies and P. Vial (1993), 'Wavelets on the interval and fast wavelet transforms', *Appl. Comput. Harm. Anal.* **1**, 54–81.

A. Cohen, I. Daubechies, O. Guleryuz and M. Orchard (1997), 'On the importance of combining wavelet-based non-linear approximation in coding strategies'. Preprint.

A. Cohen, R. DeVore and R. Hochmuth (1997), 'Restricted approximation'. Preprint.

A. Cohen, R. DeVore, P. Petrushev and H. Xu (1998), 'Nonlinear approximation and the space $\mathrm{BV}(\mathbb{R}^2)$'. Preprint.

R. R. Coifman and D. Donoho (1995) Translation invariant de-noising, in *Wavelets in Statistics* (A. Antoniadis and G. Oppenheim, eds), Springer, New York, pp. 125–150.

R. R. Coifman and M. V. Wickerhauser (1992), 'Entropy based algorithms for best basis selection', *IEEE Trans. Inform. Theory* **32**, 712–718.

S. Dahlke and R. DeVore (1997), 'Besov regularity for elliptic boundary value problems', *Commun. Partial Diff. Eqns.* **22**, 1–16.

S. Dahlke, W. Dahmen and R. DeVore (1997), Nonlinear approximation and adaptive techniques for solving elliptic operator equations, in *Multiscale Wavelet Methods for PDEs* (W. Dahmen, A. Kurdila and P. Oswald, eds), Academic Press, pp. 237–284.

W. Dahmen (1997), Wavelet and multiscale methods for operator equations, in *Acta Numerica*, Vol. 6, Cambridge University Press, pp. 55–228.

I. Daubechies (1988), 'Orthonormal bases of compactly supported wavelets', *Comm. Pure Appl. Math.* **41**, 909–996 .

I. Daubechies (1992), *Ten Lectures on Wavelets*, Vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia.

G. Davis, S. Mallat and M. Avellaneda (1997), 'Adaptive greedy approximations', *Constr. Approx.* **13**, 57–98.

R. A. DeVore (1987), 'A note on adaptive approximation', *Approx. Theory Appl.* **3**, 74–78.

R. A. DeVore and G. G. Lorentz (1993), *Constructive Approximation*, Vol. 303 of *Grundlehren*, Springer, Heidelberg.

R. A. DeVore and B. Lucier (1990), 'High order regularity for conservation laws', *Indiana Math. J.* **39**, 413–430.

R. A. DeVore and B. Lucier (1992), Wavelets, in *Acta Numerica*, Vol. 1, Cambridge University Press, pp. 1–56.

R. A. DeVore and B. Lucier (1996), 'On the size and smoothness of solutions to nonlinear hyperbolic conservation laws', *SIAM J. Math. Anal.* **27**, 684–707.

R. DeVore and V. Popov (1987), 'Free multivariate splines', *Constr. Approx.* **3**, 239–248.

R. DeVore and V. Popov (1988*a*), 'Interpolation of Besov spaces', *Trans. Amer. Math. Soc.* **305**, 397–414.

R. A. DeVore and V. A. Popov (1988*b*), Interpolation spaces and nonlinear approximation, in *Function Spaces and Applications* (M. Cwikel *et al.*, eds), Vol. 1302 of *Lecture Notes in Mathematics*, Springer, Berlin, pp. 191–205.

R. A. DeVore and K. Scherer (1979), 'Interpolation of operators on Sobolev spaces', *Ann. Math.* **109**, 583–599.

R. A. DeVore and K. Scherer (1980), Variable knot, variable degree approximation to x^β, in *Quantitative Approximation* (R. A. DeVore and K. Scherer, eds), Academic Press, New York, pp. 121–131.

R. A. DeVore and R. C. Sharpley (1984), *Maximal Functions Measuring Smoothness*, Memoirs Vol. 293, American Mathematical Society, Providence, RI.

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

R. A. DeVore and V. Temlyakov (1996), 'Some remarks on greedy algorithms', *Adv. Comput. Math.* **5**, 173–187.

R. DeVore and X. M. Yu (1986), 'Multivariate rational approximation', *Trans. Amer. Math. Soc.* **293**, 161–169.

R. DeVore and X. M. Yu (1990), 'Degree of adaptive approximation', *Math. Comput.* **55**, 625–635.

R. DeVore, R. Howard and C. A. Micchelli (1989), 'Optimal nonlinear approximation', *Manuskripta Math.* **63**, 469–478.

R. DeVore, B. Jawerth and B. Lucier (1992), 'Image compression through transform coding', *IEEE Proc. Inform. Theory* **38**, 719–746.

R. DeVore, B. Jawerth and V. Popov (1992), 'Compression of wavelet decompositions', *Amer. J. Math.* **114**, 737–785.

R. DeVore, G. Kyriazis, D. Leviatan and V. M. Tikhomirov (1993), 'Wavelet compression and nonlinear n-widths', *Adv. Comput. Math.* **1**, 197–214.

R. DeVore, B. Lucier and Z. Yang (1996), Feature extraction in digital mammography, in *Wavelets in Biology and Medicine* (A. Aldroubi and M. Unser, eds), CRC, Boca Raton, FL, pp. 145–156.

R. DeVore, W. Shao, J. Pierce, E. Kaymaz, B. Lerner and W. Campbell (1997), Using nonlinear wavelet compression to enhance image registration, in *Wavelet Applications IV, Proceedings of the SPIE Conf.* **3028**, AeroSense 97 Conference, Orlando, FL, April 22–24, 1997 (H. Szu, ed.), pp. 539–551.

R. A. DeVore, S. Konyagin and V. Temlyakov (1998), 'Hyperbolic wavelet approximation', *Constr. Approx.* **14**, 1–26.

D. Donoho (1997), 'CART and best-ortho-basis: a connection', *Ann. Statistics* **25**, 1870–1911.

D. Donoho and I. Johnstone (1994), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrika* **81**, 425–455.

D. Donoho, I. Johnstone, G. Kerkyacharian and D. Picard (1996), 'Wavelet shrinkage asymptotia?', *J. Royal Statistical Soc., Ser. B.* **57**, 301–369.

Z. Gao and R. Sharpley (1997), 'Data compression and the elementary encoding of wavelet compression'. Preprint.

E. Godlewski and P.-A. Raviart (1991), *Hyperbolic Systems of Conservation Laws*, Mathématiques et Applications, Ellipses.

A. Harten (1994), 'Adaptive multiresolution schemes for shock computations', *J. Comput. Phys.* **115**, 319–338.

D. Jerison and C. E. Kenig (1995), 'The inhomogeneous Dirichlet problem in Lipschitz domains', *J. Functional Analysis* **93**, 161–219.

R. Q. Jia (1983), Approximation by smooth bivariate splines on a three directional mesh, in *Approximation Theory IV* (C. K. Chui, L. L. Schumaker and J. Ward, eds), Academic Press, New York, pp. 539–546.

H. Johnen and K. Scherer (1977), On the equivalence of the K-functional and moduli of continuity and some applications, in *Constructive Theory of Functions of Several Variables*, Vol. 571 of *Lecture Notes in Mathematics*, Springer, Berlin, pp. 119–140.

L. Jones (1992), 'A simple lemma on greedy approximation in Hilbert space and convergence results for projection pursuit regression and neural network training', *Ann. Statistics* **20**, 608–613.

J. P. Kahane (1961), *Teoria Constructiva de Functiones*, Course notes, University of Buenos Aires.

B. Kashin (1977) 'The widths of certain finite dimensional sets and classes of smooth functions', *Izvestia* **41**, 334–351.

B. Kashin and V. Temlyakov (1997), 'On best n-term approximation and the entropy of sets in the space L_1', *Math. Notes* **56**, 1137–1157.

V. A. Kondrat'ev and O. A. Oleinik (1983), 'Boundary value problems for partial differential equations in non-smooth domains', *Russian Math. Surveys* **38**, 1–86.

G. Kyriazis (1996) 'Wavelet coefficients measuring smoothness in $H_p(\mathbb{R}^d)$', *Appl. Comput. Harm. Anal.* **3**, 100–119

B. Lucier (1986), 'Regularity through approximation for scalar conservation laws', *SIAM J. Math. Anal.* **19**, 763–773.

G. G. Lorentz, M. von Golitschek and Ju. Makovoz (1996), *Constructive Approximation: Advanced Problems*, Springer, Berlin.

M. McClure and L. Carin (1997), 'Matched pursuits with a wave-based dictionary'. Preprint.

V. Mairov and J. Ratasby (1998), 'On the degree of approximation using manifolds of finite pseudo-dimension'. Preprint.

S. Mallat (1989), 'Multiresolution and wavelet orthonormal bases in $L_2(\mathbb{R})$', *Trans. Amer. Math. Soc.* **315**, 69–87.

S. Mallat (1998), *A Wavelet Tour of Signal Processing*, Academic Press, New York.

S. Mallat and F. Falzon (1997), 'Analysis of low bit rate image transform coding'. Preprint.

Y. Meyer (1990) *Ondelettes et Opérateurs, Vols 1 and 2*, Hermann, Paris.

D. Newman (1964), 'Rational approximation to $|x|$', *Michigan Math. J.* **11**, 11–14.

E. Novak (1996), 'On the power of adaptation', *J. Complexity* **12**, 199–237.

K. Oskolkov (1979), 'Polygonal approximation of functions of two variables', *Math. USSR Sbornik* **35**, 851–861.

P. Oswald (1980), 'On the degree of nonlinear spline approximation in Besov–Sobolev spaces', *J. Approx. Theory* **61**, 131–157.

J. Peetre (1963) *A Theory of Interpolation of Normed Spaces*, Course notes, University of Brasilia.

A. Pekarski (1986), 'Relations between the rational and piecewise polynomial approximations', *Izvestia BSSR, Ser. Mat.-Fiz. Nauk* **5**, 36–39.

V. Peller (1980), 'Hankel operators of the class S_p, investigations of the rate of rational approximation and other applications', *Mat. Sbornik* **122**, 481–510.

P. P. Petrushev (1986), 'Relations between rational and spline approximations in L_p metrics', *J. Approx. Theory* **50**, 141–159.

P. P. Petrushev (1988), Direct and converse theorems for spline and rational approximation and Besov spaces, in *Function Spaces and Applications* (M. Cwikel *et al.*, eds), Vol. 1302 of *Lecture Notes in Mathematics*, Springer, Berlin, pp. 363–377.

P. Petrushev and V. Popov (1987), *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge.

G. Pisier (1980) 'Remarques sur un résultat non publié de B. Maurey', *Seminaire d'Analyse Fonctionelle 1980–81, École Polytechnique, Centre de Mathématiques, Palaiseau*.

E. Schmidt (1907), 'Zur Theorie der linearen und nichtlinearen Integralgleichungen. I', *Math. Ann.* **63**, 433–476.

I. Schoenberg (1946), 'Contributions to the problem of approximation of equidistant data by analytic functions', *Quart. Appl. Math.* **4**, 45–99.

J. Shapiro (1993), An embedded hierarchial image coder using zerotrees of wavelet coefficients, in *Data Compression Conference* (J. A. Storer and M. Cohn, eds), IEEE Computer Society Press, Los Alamitos, CA, pp. 214–223.

E. Stein (1970), *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton.

G. Strang and G. Fix (1973), A Fourier analysis of the finite element variational method, in *Constructive Aspects of Functional Analysis* (G. Geymonat, ed.), C.I.M.E., II, Ciclo, 1971, pp. 793–840.

V. Temlyakov (1998*a*), 'Best m-term approximation and greedy algorithms'. Preprint.

V. Temlyakov (1998*b*), 'Nonlinear m-term approximation with regard to the multivariate Haar system'. Preprint.

J. F. Traub, G. W. Wasilkowski and H. Woźniakowski (1988), *Information-Based Complexity*, Academic Press, Boston.

V. N. Vapnik (1982), *Estimation of Dependences Based on Empirical Data*, Springer, Berlin.

M. V. Wickerhauser (1994), *Adapted Wavelet Analysis from Theory to Software*, Peters.

Z. Xiong, K. Ramchandran and M. T. Orchard (1997), 'Space-frequency quantization for wavelet image coding', *Trans. Image Processing* **6**, 677–693.