Course Notes Bilbao 2013
Numerical Methods in High Dimensions
Lecture 1: The Role of Approximation Theory

Ronald DeVore

April 11, 2013

Abstract

In this first lecture, we give the setting of classical approximation theory and mention its drawbacks when the approximation takes place in high dimensions.

1 Introduction and Motivation

Many application domains require some form of approximation in high dimensions. By this, we mean that the function u, which we wish to capture numerically, depends on many variables or parameters. In these lectures, we will primarily focus on the following two examples:

Sensing/active learning:
In this setting, we may ask for the values $\lambda_j(u)$, $j = 1, \ldots, n$, for *any* linear functionals λ_j. The main issue is to determine for a given n the best question to ask, i.e. the best functionals to choose, in order to approximate u effectively using the drawn information. Sometimes the form of the linear functionals is restricted so that they can be implemented in practice. Point evaluation queries is a very common restriction. The field of compressed sensing falls into this paradigm when one considers the discrete setting and will be the subject of our second lecture. In our third lecture, we will consider the case of recovering a function u of many variables.

Operator equations:
Frequently, we are not given direct access to u, but only know this function as a solution to some (linear or nonlinear) equation $F(u) = 0$. A typical example is $Au = f$ where A is a known bounded mapping on some relevant spaces. Partial differential and integral equations fall into this category. We have complete knowledge of the operator A, the function f and any additional information (such as boundary or initial conditions) that are needed to uniquely determine u. We are given a computational budget n and we wish to approximate u by a function $\tilde{u}$ as efficiently as possible within this budget. For example, n could represent the number of floating point operations or memory calls.

To begin the discussion and focus in on the problem to be circumvented, we first consider what one classically does in approximation theory. This will expose the so-called 'curse of dimensionality', which, in a nutshell, says that we cannot effectively treat such high dimensional problems without some new ideas. We will then discuss, in later lectures, some of the new ideas currently being developed for high dimensional problems.

2 The canonical setting for approximation

The goal of approximation theory is to approximate a given function u (called the *target function*) by simpler, easier to compute functions, such as polynomials, rational functions, or piecewise polynomials[1] It has a long history beginning with the famous Bolzano-Weierstrass theorem (1817) which says that each continuous function can be approximated uniformly with arbitrary accuracy by algebraic polynomials. The field of approximation theory emerged to aswer the rate-distortion question: how fast does the approximation error decay as the degree of the polynomial increases? Of course, this rate-distortion depends on the target function. But which properties of u govern this rate?

2.1 The metric

A precise formulation of approximation requires us to specify how we will measure the error (or distortion) and what are the functions that we wish to approximate. This is done by providing a space X which contains the set $\mathcal{K}$ of functions we wish to approximate and a norm $\|\cdot\|_X$ defined on X to measure the distance

$$d(f,g)=\|f-g\|_X, \quad f,g\in X, \tag{2.1}$$

which will be used to measure distortion. In most cases of interest, X is a linear space which is complete with respect to this norm and is therefore a Banach space.

The most commonly used distortion metrics are the L_p norms. If (Ω,Σ,μ) is a measurable space, then $L_p(\Omega,\mu)$ denotes the set of all measurable functions f such that the quantity

$$\|u\|_{L_p(\Omega,\mu)} := \begin{cases} (\int\limits_{\Omega} |u(x)|^p \, d\mu(x))^{1/p}, & 0\le p<\infty, \\ \operatorname*{esssup}\limits_{x\in\Omega} |f(x)|, & p=\infty, \end{cases} \tag{2.2}$$

is finite. These are norms when $p\ge 1$, and quasi-norms when $0<p<1$. In most applications, $d\mu$ is Lebesgue measure. In this case we simply write dx and $L_p(\Omega)$.

A common choice of metric for the **Sensing Problems** is an L_2 norm. This is especially the case when the measurements are noisy, since then the averaging in the L_2 norm helps avoid fitting the noise. However, in some applications an L_∞ or L_1 norm may be an appropriate choice. For **Operator Equations**, the choice of norms is more subtle and depends on finding the setting in which uniqueness and well posedness holds. For example, in second order elliptic problems the typical choice is the space H^1 whose semi-norm is the sum of the L_2 norms of the first derivatives of u (see §4 where smoothness spaces are defined).

[1] Sometimes, we will denote the target function by f as is more customary in approximation theory.

2.2 The approximation tool

Once, we have settled on the distortion norm, the next step is to specify which functions will be used for the approximation; these functions are called the *approximation tool.* The historical development of approximation theory, which took place at the beginning of the last century, concentrated on using algebraic or trigonometric polynomials for the approximation process. One gains accuracy in the approximation by increasing the degree of the polynomials. The reader has certainly been exposed to Fourier series whose partial sums are trigonometric polynomials that are often used for approximation and analysis.

With the advent of modern computers, it became clear that polynomials were cumbersome and often unstable in computation, when their degree is large. This led to the development of piecewise polynomials and spline functions (piecewise polynomials with prescribed smoothness where the pieces join), starting with the work of Schoenberg [31] in the 1940's. The idea was to keep the degree of the polynomial pieces small but increase approximation efficiency by partitioning the domain of u into many small cells. The univariate theory of splines is simple and complete (see [2, 32]). The multivariate theory is complicated when global smoothness is required of the piecewise polynomial and the most penetrating results fall into the domain of finite elements (see [9, 8, 10]).

Beyond piecewise polynomials, a myriad of novel approximation tools were developed, often directed at specific application domains. Among these, the most notable developments are

Wavelets: These were initially developed for image and signal processing but later applied with much success to operator equations (see [19, 11, 24]).

Radial Basis Functions: These are often applied to data fitting, especially when fitting functions of many variables [34]).

Rational functions and Padé expansions: These are especially important when treating functions of complex variables (see [3]).

Dictionaries: These are used to add flexibility to the choice of a basis or representation system when doing approximation (see [33]).

2.3 Linear or nonlinear

One can broadly divide approximation tools into two classes: linear and nonlinear. In linear approximation, the approximation process takes place from a sequence of **linear** spaces X_n, $n = 1, 2, \ldots$. By using the space $X_0 := \{0\}$ and, if necessary, repeating the spaces in this sequence, we can assume $\dim(X_n) = n$. Increasing n results in improved accuracy in the approximation.

Spurred on by application demands, there has been an emergence of approximation methods, known as *nonlinear approximation.* In the nonlinear setting, the linear space X_n is replaced by a nonlinear set Σ_n which can be described by $O(n)$ parameters. A simple but important example is to take for Σ_n the collection of all piecewise constant functions on $[0, 1]$ which consist of n pieces. Adding two elements of Σ_n together generally does not land in Σ_n but rather in Σ_{2n}.

The class of rational functions of degree n is another example of a nonlinear set Σ_n that is used in several application domains. A general nonlinear procedure that is often used in practice is to take a basis, e.g. a wavelet basis, or more generally a dictionary $\mathcal{D}$ of functions, and define Σ_n as the set of all functions which are a linear combination of n elements from $\mathcal{D}$. In this case $\Sigma_n + \Sigma_n = \Sigma_{2n}$. Nonlinear methods of approximation offer a distinct advantage over linear methods in approximation accuracy, as will be explained in the following sections.

3 Approximation spaces

Once the approximation tool is chosen, the goal of approximation theory is to give an exact characterization of which functions have a given rate of approximation. To describe this pursuit, we consider the standard setting in linear approximation. We start with a Banach space X and the distortion norm $\|\cdot\|_X$. For the approximation tool, we take a sequence $(X_n)_{n\geq 1}$ of linear spaces, $X_n \subset X$ with $\dim(X_n) = n$. For a given $u \in X$, the error of approximation

$$e_n(u) := e_n(u)_X = e(u, X_n)_X := \inf_{g\in X_n} \|u - g\|_X, \tag{3.1}$$

tells us the optimal performance we can obtain when using the elements of X_n for the approximation of u. A function $u_n \in X_n$ for which $\|u - u_n\|_X = e_n(u)$ is called a best approximation. A compactness argument shows that best approximations always exist but they need not be unique, and even when unique are difficult to find. If $\bigcup_{n\geq 1} X_n$ is dense in X, then $e_n(u)$ will tend to zero as n tends to infinity. How fast it tends to zero for a specific u gives a quantitative description of how effective this approximation process is for resolving u.

An approximation method, or a numerical procedure that uses the spaces $(X_n)_{n\geq 1}$, is said to be *instance optimal* if it produces for each n an approximation $\hat{u}_n \in X_n$ that satisfies

$$\|u - \hat{u}_n\|_X \leq C e_n(u)_X, \tag{3.2}$$

with an absolute constant C that is in dependent of u and n. In most application settings, it is rare to achieve instance optimality and so one introduces weaker measures of performance.

The most popular of these is the notion of approximation classes. Given a real number $r > 0$, we define $\mathcal{A}^r := \mathcal{A}^r(X, (X_n)_{n\geq 1})$ as the set of all $u \in X$ for which

$$e_n(u)_X \leq M n^{-r}, \quad n = 1, 2, \ldots, \tag{3.3}$$

and the smallest value of M is defined as the semi-norm $|u|_{\mathcal{A}^r}$ on $\mathcal{A}^r$. One could equally well replace the sequence $(n^{-r})_{n\geq 1}$ by a more general decreasing sequence $(\epsilon_n)_{n\geq 1}$ but for most numerical applications the spaces $\mathcal{A}^r$ suffice.

Understanding the space $\mathcal{A}^r$ is important in numerical analysis because it tells us the performance (in terms of rate of approximation) we can expect of a numerical algorithm built on a chosen approximation tool. This, in turn, tells us the accuracy we can achieve in the approximation for a given computational budget.

4 Smoothness spaces

The fundamental problem in approximation theory is to give an intrinsic characterization of the approximation classes $\mathcal{A}^r(X,(X_n))$. This problem has occupied approximation theory for decades and it has not only been resolved for all of the standard approximation methods, but also certain unifying principles have emerged which direct one how to proceed in any new setting. The characterization of $\mathcal{A}^r$ is almost always in terms of smoothness spaces, most prominently the Besov spaces.

We briefly recall the standard hierarchy of smoothness spaces. The simplest of these are the Sobolev spaces $W^k(L_p(\Omega))$, $1 \le p \le \infty$, $k = 1, 2, \ldots$, defined on a domain $\Omega \subset \mathbb{R}^d$. This space consists of all functions whose (weak) derivatives of order k are all in $L_p(\Omega)$. The semi-norm $|\cdot|_{W^k(L_p(\Omega))}$ is given by

$$|u|_{W^k(L_p(\Omega))} := \sum_{|\alpha|=k} \|D^\alpha u\|_{L_p(\Omega)}. \tag{4.1}$$

When $p = \infty$, $W^k(L_\infty(\Omega))$ is often replaced by $C^k(\Omega)$ which adds the stipulation that the derivatives of order k should all be continuous. We obtain the norm on $W^k(L_p(\Omega))$ by adding the $L_p(\Omega)$ norm to the semi-norm

$$\|u\|_{W^k(L_p(\Omega))} := |u|_{W^k(L_p(\Omega))} + \|u\|_{L_p(\Omega)}. \tag{4.2}$$

Lipschitz and Besov spaces generalize the Sobolev spaces to fractional order $s > 0$ and also cover the full range $0 < p \le \infty$. They are a staple in modern analysis and found in many text books (see e.g. [1, 23, 4]). For $0 < s \le 1$, a continuous function u is said to be Lipschitz continuous of order s ($u \in \operatorname{Lip} s$) if

$$|u(x+h) - u(x)| \le M|h|^s, \quad x, x+h \in \Omega, \tag{4.3}$$

where $M > 0$ is independent of x and h. This space is a Banach space equiped with the norm

$$\|u\|_{\operatorname{Lip} s} := \|u\|_{L^\infty(\Omega)} + |u|_{\operatorname{Lip} s}, \tag{4.4}$$

where the the seminorm is defined as the smallest constant M such that (4.3) holds.

To generalize this definition to larger values of s, and to functions in $L_p(\Omega)$, $0 < p < \infty$, we use higher order differences, defined for integers $m \ge 1$ by

$$\Delta_h^m(u,x) := (-1)^m \sum_{k=0}^{m} (-1)^k \binom{m}{k} u(x+kh), \quad h \in \mathbb{R}^d. \tag{4.5}$$

Here, we use the convention that this difference is defined to be zero if any of the points $x, x+h, \ldots, x+mh$ lie outside the domain Ω on which u is defined. Thus, $\Delta_h(u,x) := u(x+h) - f(x)$ and the higher order difference Δ^m is the m-fold composition of Δ.

From these differences, we can define the moduli of smoothness

$$\omega_m(u,t)_{L_p(\Omega)} := \sup_{|h|\le t} \|\Delta_h^m(u,\cdot)\|_{L_p(\Omega)}, \quad t > 0. \tag{4.6}$$

For each $u \in L_p(\Omega)$, the modulus $\omega_m(u,t)_{L_p(\Omega)}$ tends to zero when $t \to 0$. The speed of this decay tells us how smooth u is.

The Besov spaces classify the smoothness of a function in $L_p(\Omega)$ by the speed at which $\omega_m(u,t)$ tends to zero. The simplest of these Besov spaces are the $B^s_\infty(L_p(\Omega))$ spaces which are defined as the set of all $u \in L_p(\Omega)$ for which

$$\omega_m(u,t)_{L_p(\Omega)} \leq Mt^s, \quad t > 0, \tag{4.7}$$

where $m := [s] + 1$ (it turns out that any $m > s$ gives the same space with an equivalent norm). The smallest M for which (4.7) holds is the semi-norm $|u|_{B^s_\infty(L_p(\Omega))}$ and

$$\|u\|_{B^s_\infty(L_p(\Omega))} := |u|_{B^s_\infty(L_p(\Omega))} + \|u\|_{L_p(\Omega)}. \tag{4.8}$$

Notice that these spaces are defined for all $s > 0$ and $0 < p \leq \infty$. The reader should think of these spaces as a way to define analogues of $W^s(L_p(\Omega))$ for all $s > 0$, $0 < p \leq \infty$. However, note that the Besov and Sobolev spaces do not necessarilly coincide where they are both defined.

The general Besov space $B^s_q(L_p(\Omega))$ has in addition to s and p a third index q which is a fine tuning parameter. The semi-norm on this space is defined by

$$\|u\|_{B^s_q(L_p(\Omega))} := \{\int\limits_0^\infty [t^{-s}\omega_m(u,t)_{L_p(\Omega)}]^q \, \frac{dt}{t}\}^{1/q}. \tag{4.9}$$

Decreasing the value of q corresponds to a slightly stronger requirement placed on u.

The reader should not be overwhelmed by the definition of the Besov spaces. Think of $B^s_q(L_p(\Omega)$ as a space of functions with s derivatives in L_p. The actual definitions are complicated because we want to allow fractional s and values of p smaller than one.

5 Characterization of approximation spaces

There is a simple unifying principle for characterizing $\mathcal{A}^r$. It establishes two inequalities that link the approximation process to smoothness spaces. To illustrate this procedure, we consider a very classical setting in which the space $X = C[-\pi,\pi]$ consists of continuous and 2π periodic functions equipped with the L_∞ norm, and the approximation tool consists of the spaces $X_{2n+1} = \mathbf{T}_n$ of trigonometric polynomials of degree $\leq n$. This is a linear space of dimension $2n+1$ with a basis given by $1, \cos x, \sin x, \ldots, \cos nx, \sin nx$.

The most efficient way to characterize the approximation classes $\mathcal{A}^r$ for trigonometric approximation is to prove the following two inequalities which hold for all positive integers k.

Jackson inequality: For each continuous periodic function $u \in W^k(L_\infty[-\pi,\pi]$, we have

$$e_n(u) \leq C_k \|u^{(k)}\|_{L_\infty} n^{-k} = C_k n^{-k} |u|_{W^k(L_\infty[-\pi,\pi])}, \quad n = 1, 2, \ldots. \tag{5.1}$$

Bernstein inequality: For each $T \in \underline{\mathrm{T}}_n$,

$$|T|_{W^k(L_\infty[-\pi,\pi])} = \|T^{(k)}\|_{C[-\pi,\pi]} \leq n^k \|T\|_{C[-\pi,\pi]}, \quad n = 1, 2, \ldots. \tag{5.2}$$

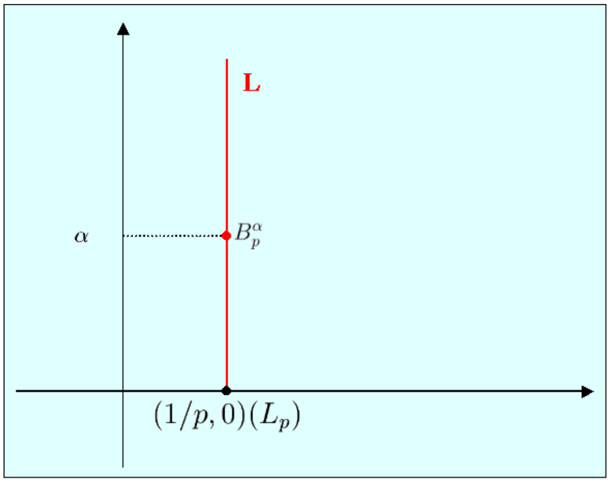

Figure 5.1: Graphical Depiction of Smoothness Spaces for Linear Approximation. The approximation space $\mathcal{A}_\infty^\alpha((X_n), L_p)$ coincides with the Besov space $B_p^\alpha = B_\infty^\alpha(L_p(\Omega))$.

The first inequality is named after Dunham Jackson, who actually proved more subtle inequalities for trigonometric approximation that include (5.1). The second is commonly attributed to Sergei Bernstein. Notice that these inequalities pair together in that the semi-norm $|u|_{W^k(L_\infty)}$ and the factor n^k appearing in both inequalities is the same. For other approximation tools, the challenge is to find a corresponding pair of Jackson and Bernstein inequalities. A typical setting is when linear approximation is done in $L_p(\Omega)$ in which case $W^k(L_\infty(\Omega))$ is replaced by $W^k(L_p(\Omega))$.

Once a pair of Jackson and Bernstein inequalities have been found, the task of characterizing the approximation classes $\mathcal{A}^r$ can be completed by using results from the interpolation of linear operators (see e.g. Chapter 7 of [23]). In the case under discussion, we have that for any $0 < r < k$, $\mathcal{A}^r$ coincides with the real interpolation space $(C[-\pi,\pi], C^k[-\pi,\pi])_{r/k,\infty}$ with equivalent norms. Space does not allow us to go into a lengthy discussion of these interpolation spaces but this is covered in standard textbooks, e.g. [4, 5]. The good news is that these interpolation spaces are known, not only for the pair C, C^k but for most other pairs of classical spaces that appear when analyzing approximation processes. In particular, it is known that $(C[-\pi,\pi], C^k[-\pi,\pi])_{r/k,\infty}$ coincides with the Besov space $B_\infty^r(L_\infty[-\pi,\pi])$, thereby showing that $\mathcal{A}^r$ is also equal to this Besov space with equivalent norms. A similar result holds in the case the approximation is done in $L_p(\Omega)$. Figure 1 gives a graphical description of the above fact. We use a point in the upper right quadrant of $\mathbb{R}^2$ to represent a smoothness space. Such a point (x, y) has two coordinates. The first coordinate depicts the L_p space through $x = 1/p$ and the second coordinate depicts the smoothness $y = r$.

So, characterizing the approximation spaces $\mathcal{A}^r$ for a given approximation tool is reduced to deriving the correct Jackson and Bernstein inequalities for this tool. This may be subtle. In the case of approximation by trigonometric polynomials, we have noted that this follows from classical results on approximation. Let us note that to establish the Jackson inequality in the above setting of trigonometric approximation, one is tempted to use the n-th partial sum $S_n(f)$

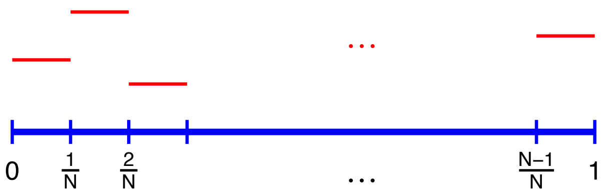

Figure 6.2: The graph of a typical function f in $\mathcal{S}_n$.

of the Fourier series since obviously $e_n(u) \leq \|u - S_n(u)\|_{C[-\pi,\pi]}$. But this does not quite work since the operators S_n are not even uniformly bounded on $C[-\pi, \pi]$. Instead, one has to modify $S_n(u)$ through summability methods such as the Cesaro sums. A general summability that works for all $k = 1, 2, \ldots$ was given by Jackson.

6 Nonlinear Approximation

We have already mentioned several examples of nonlinear approximation. To draw out the advantages of nonlinear approximation over its linear counterpart, we consider the simple case of piecewise constant approximation on the interval $\Omega := [0, 1]$. Many numerical algorithms, including the prominent finite element methods, are built on approximation by piecewise polynomials over partitions of a domain. The popularity of piecewise polynomials stems from the fact that they are among the simplest functions to compute numerically. While piecewise constant approximation on an interval is, albeit, very simple, most of the results we discuss have analogues for more general polynomial degree and can also be extended to higher Euclidean space dimension. To do our comparison between linear and nonlinear approximation, we measure distortion in the L_∞ norm. Similar results hold for approximation in L_p norms or even in Sobolev or Besov norms.

6.1 Linear approximation with piecewise constants

In linear approximation, the partitions are fixed in advance. The most natural and democratic choice is to choose intervals of equal size. We therefore introduce for every integer $n \geq 1$, the uniform partition

$$\mathcal{P}_n := \left\{ \Big[\frac{k-1}{n}, \frac{k}{n}\Big) \ : \ k = 1, \ldots, n \right\}. \tag{6.1}$$

We denote by $\mathcal{S}_n$ the linear space consisting of all piecewise constant functions subordinate to $\mathcal{P}_n$, i.e., all functions of the form

$$g = \sum_{I \in \mathcal{P}_n} c_I \chi_I, \tag{6.2}$$

where c_I are real numbers and χ_A denotes the characteristic function of a set A. Thus, $\mathcal{S}_n$ is a linear space of dimension n spanned by the set $\{\chi_I\}_{I \in \mathcal{P}_n}$ of basis elements. Figure 6.1 depicts a typical function in $\mathcal{S}_n$

The spaces $\mathcal{A}^r(C(\Omega),(\mathcal{S}_n)_{n\geq 1})$ can be characterized in terms of a very simple notion of smoothness. The characterization of approximation spaces referred to above is given by

$$\mathcal{A}^r = \mathcal{A}^r(C(\Omega),(\mathcal{S}_n)_{n\geq 1}) = \operatorname{Lip} r, \quad 0 < r \leq 1, \tag{6.3}$$

and the $\mathcal{A}^r$ norm is equivalent to the $\operatorname{Lip} r$ norm[2]

$$\|u\|_{\operatorname{Lip} r} \lesssim \|u\|_{\mathcal{A}^r} \lesssim \|u\|_{\operatorname{Lip} r}, \tag{6.4}$$

with absolute constants of equivalence for each r. The limitation $r \leq 1$ in (6.4) is natural when considering the process of approximation by piecewise constants since it is easy to prove that a function u such that $e_n(u) \leq Cn^{-r}$ for some $r > 1$ is necessarily constant. If we would instead approximate by piecewise polynomials of degree m then the range of r for which the approximation spaces are non-trivial would be $0 < r \leq m+1$.

While we have advertised that the general way to characterize the approximation classes for a given approximation tool is to prove Jackson and Bernstein inequalities, in the simple setting of piecewise constant approximation, it is easy to prove the characterization directly and we would be remiss not to point this out. For example, to prove the upper inequality in (6.4), for $u \in C(\Omega)$, we define

$$g := \sum_{I\in\mathcal{P}_n} u(\xi_I)\chi_I, \tag{6.5}$$

where ξ_I is the midpoint of I. We clearly have for any $x \in I$ that

$$|u(x) - u(\xi_I)| \leq |u|_{\operatorname{Lip} r}(|I|/2)^r = |u|_{\operatorname{Lip} r}(2n)^{-r}, \tag{6.6}$$

and therefore

$$e_n(u) \leq \|u-g\|_{L^\infty(\Omega)} \leq |u|_{\operatorname{Lip} r} n^{-r}. \tag{6.7}$$

This gives $|u|_{\mathcal{A}^r} \leq |u|_{\operatorname{Lip} r}$, and therefore the upper inequality in (6.4) holds with constant 1.

To prove the lower inequality, let $u \in \mathcal{A}^r$ and fix any two points $x, y \in \Omega$. We choose the integer $n \geq 2$ such that $\frac{1}{n} < |x-y| \leq \frac{1}{n-1}$. Let $g \in \mathcal{S}_n$ satisfy $\|u-g\|_{L_\infty(\Omega)} \leq |u|_{\mathcal{A}^r} n^{-r}$. If x', y' are two points that lie in the same interval $I \in \mathcal{P}_n$, we have $g(x') = g(y')$ and therefore,

$$|u(x') - u(y')| \leq |u(x') - g(x')| + |u(y') - g(y')| \leq 2|u|_{\mathcal{A}^r} n^{-r} \leq 2|u|_{\mathcal{A}^r}|x-y|^r. \tag{6.8}$$

Returning to our points x, y, we have $x \in I$ and $y \in I'$ with $I, I' \in \mathcal{P}_n$ and the intervals I, I' share a common endpoint a. Applying (6.8) to x, a and then to y, a we derive that

$$|u(x) - u(y)| \leq |u(x) - u(a)| + |u(y) - u(a)| \leq 4|u|_{\mathcal{A}^r}|x-y|^r. \tag{6.9}$$

Hence $|u|_{\operatorname{Lip} r} \leq 4|u|_{\mathcal{A}^r}$ and we arrive at the lower inequality in (6.4) with constant 1/4.

The right inequality in (6.4) is called a *direct theorem.* It shows that the smoothness condition $u \in \operatorname{Lip} r$ guarantees the rate of approximation $\mathcal{O}(n^{-r})$ by the elements of $\mathcal{S}_n$. The left inequality in (6.4) shows that a rate of approximation implies that u must have a certain smoothness. Such an implication is called an *inverse theorem.*

We see that the above result conforms to our general depiction in Figure 5.1 of approximation spaces being characterized by Besov spaces for linear methods of approximation.

[2] We use the notation $A \lesssim B$ to mean that $A \leq CB$ with a constant independent of the primary parameters on which A and B belong. When necessary we indicate the dependence of C on parameters.

6.2 Nonlinear approximation

Now let us turn to nonlinear approximation by piecewise constants in order to see the gain in approximation power. In nonlinear approximation, we do not constrain the partitions but only the number of cells in such a partion. Accordingly, we define Σ_n as the set of all piecewise constant functions with at most n pieces, namely, the set of all functions of the form

$$g := \sum_{I \in \mathcal{P}} c_I \chi_I, \tag{6.10}$$

where $\mathcal{P}$ is **any** partition of Ω such that $\#(\mathcal{P}) \leq n$. Clearly, Σ_n is not a linear space since the partition $\mathcal{P}$ can change at each instance. Note, however, that we still have $\Sigma_n + \Sigma_n \subset \Sigma_{2n}$.

We define the approximation error

$$\sigma_n(u) := \sigma_n(u)_{C(\Omega)} := \inf_{g \in \Sigma_n} \|u - g\|_{L_\infty(\Omega)}, \tag{6.11}$$

and define the the approximation classes $\mathcal{A}^r(C(\Omega), (\Sigma_n)_{n\geq 1})$ by using σ_n in place of e_n in (3.3).

To uncover the gain in nonlinear approximation, it is sufficient to consider the case $r = 1$ where we have the following result of Kahane [27]:

$$\mathcal{A}^1(C(\Omega), (\Sigma_n)_{n\geq 1}) = BV(\Omega) \cap C(\Omega), \tag{6.12}$$

and, moreover,

$$2|u|_{\mathcal{A}^1(C(\Omega),(\Sigma_n)_{n\geq 1})} = V(u, \Omega). \tag{6.13}$$

Here, for any interval I, $V(u, I)$ is the variation of u on I, which is defined as

$$V(u, I) := \sup \sum_{i=1}^{m} |u(x_i) - u(x_{i-1})|, \tag{6.14}$$

with the supremum taken over all finite sequences $x_0 < x_1 < \cdots < x_m$ contained in I.

We give a proof of (6.13) since it illustrates some overiding principles that are used in nonlinear approximation. The first is the principle of error equidistribution which is used when establishing direct theorems. It begins with a surrogate for the error of approximation by constants on an interval I. Given such an interval, the median value m_I of u on I satisfies

$$\|u - m_I\|_{L_\infty(I)} \leq V(u, I)/2. \tag{6.15}$$

The principle of error equidistribution says we should choose our partition $\mathcal{P}^*$ into n intervals so as to equalize the surrogate bound, i.e. $V(u, I)$ should be the same for each $I \in \mathcal{P}^*$.

The variation $V(u, [a, b])$ of a continuous function u on the interval $[a, b]$ depends continuously on a, b and is set-additive in the sense that $V(u, [a, b]) = V(u, [a, c]) + V(u, [c, b])$ for $a < c < b$. Thus, given our budget n, we can choose points $0 =: x_0 < x_1 < \cdots < x_n := 1$ such that

$$V(u, I_k) = \frac{M}{n}, \quad k = 1, \ldots, n, \tag{6.16}$$

where $I_k := [x_{k-1}, x_k)$ and $M := V(u, \Omega)$. Hence the function $g := \sum_{k=1}^{n} m_{I_k} \chi_{I_k}$ is in Σ_n and satisfies

$$\sigma_n(u) \leq \|u - g\|_{L^\infty(\Omega)} \leq \frac{V(u, \Omega)}{2n}. \tag{6.17}$$

This proves that $2|u|_{\mathcal{A}^1(C(\Omega),(\Sigma_n),_{n\geq 1}} \leq V(u,\Omega)$.

We now consider the converse direction. We assume that $u \in \mathcal{A}^1(C(\Omega),(\Sigma_n)_{n\geq 1})$ and estimate its variation. Let $g_n \in \Sigma_n$ satisfy

$$\|u - g_n\|_{L_\infty(\Omega)} \leq M/n, \quad n = 1, 2, \ldots \tag{6.18}$$

with $M := |u|_{\mathcal{A}^1}$. We first estimate the variation of g_n for any fixed n. Let $0 = x_0(n) < x_1(n) < \cdots < x_n(n)$ be the breakpoints of the partition for g_n. Since u is continuous,

$$\begin{aligned} V(g_n,\Omega) &= \sum_{j=1}^{n-1} |g_n(x_j(n)+) - g_n(x_j(n)-)| \\ &\leq \sum_{j=1}^{n-1} |g_n(x_j(n)+) - u(x_j(n)+)| + |g_n(x_j(n)-) - u(x_j(n)-)| \\ &\leq (n-1)\frac{2M}{n} \leq 2M. \end{aligned} \tag{6.19}$$

Now, given any points $t_0 < t_1 < \cdots < t_k$ from Ω, we have

$$\sum_{j=1}^{k} |u(t_j) - u(t_{j-1})| = \lim_{n\to\infty} \sum_{j=1}^{k} |g_n(t_j) - g_n(t_{j-1})| \leq \lim_{n\to\infty} V(g_n,\Omega) \leq 2M. \tag{6.20}$$

We thereby conclude that $V(f,\Omega) \leq 2|u|_{\mathcal{A}^1}$, as desired.

Let us clearly understand the gain in nonlinear approximation in this example. The nonlinearity allows us to establish an error bound $\sigma_n(u) \leq C/n$ by only assuming u has bounded variation. This is a much weaker smoothness assumption than $u \in \mathrm{Lip}\,1$ which was needed to guarantee the same rate for the linear approximation error $e_n(u)$ according to (6.3). For example, if $u' \in L^1(\Omega)$ then u has bounded variation and is continuous. On the other hand, $u \in \mathrm{Lip}\,1$ is equivalent to $u' \in L^\infty(\Omega)$. So we have the same approximation rate in nonlinear approximation for a much broader class of functions. This is depicted in Figure 6.3. We see that the approximation space $\mathcal{A}^\alpha_q((\Sigma_n), L_p(\Omega))$ corresponds to a smoothness space on the Sobolev embedding line for embedding into L_p. What we do not gain in nonlinear approximation is a better convergence rate for functions which are already in $\mathrm{Lip}\,1$: this remain $O(1/n)$ whether we use linear or nonlinear methods.

Another important observation is that for an individual function u, the rate of decay of $\sigma_n(u)$ might be significantly faster than that of $e_n(u)$. Consider for example, the function $u(t) = t^s$ for some $0 < s < 1$. Such a function belongs to $\mathrm{Lip}\,r$ if and only if $r \leq s$ and therefore the rate of decay of $e_n(u)$ is exactly n^{-s}. On the other hand, this function has bounded variation and therefore the rate of decay of $\sigma_n(u)$ is n^{-1}.

6.3 Numerical considerations

Fnding a partition that exactly achieves equidistribution of the local approximation error is generally not numerically implementable. So, other approaches are needed for a reasonable

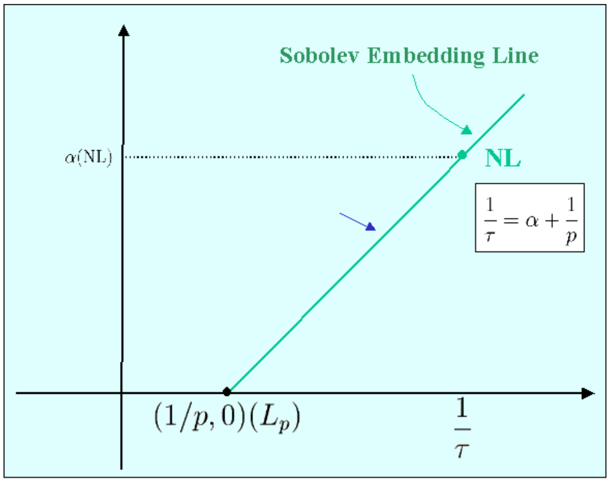

Figure 6.3: The figure depicts the smoothness spaces corresponding to nonlinear approximation. Notice that these spaces lie on the Sobolev embedding line for embedding of smoothness spaces into L_p.

numerical implementation of nonlinear approximation by piecewise constants. The most popular alternative is to use some form of *adaptive refinement*. We shall see that these methods give nearly the same performance as equidistributed partitions.

We fix $u \in C(\Omega)$ and any prescribed target accuracy $\varepsilon > 0$ and describe an *adaptive algorithm* that constructs a partition $\mathcal{P}_\varepsilon$ such that the error ε is achieved by piecewise constant approximation on $\mathcal{P}_\varepsilon$. The partition is generated by a sequence of refinements based on dyadic splits. For each interval I, we denote by $e(u, I)$, the error in approximating u by a constant in the uniform norm $\| \cdot \|_{C(I)}$ on I. It is easy to check that $2(u, I) = M_I^+ - M_I^-$, where M_I^+ and M_I^- are the maximum and minimum values taken by u on I. The algorithm begins by checking whether $e(u, \Omega) \le \varepsilon$. If it is, the algorithm stops and takes $\mathcal{P}_\varepsilon := \{\Omega\}$. If $e(u, \Omega) > \varepsilon$, then Ω is split into its two half-intervals $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$, called *children*. The current partition is then $\mathcal{P} = \{[0, \frac{1}{2}), [\frac{1}{2}, 1)\}$, i.e., Ω has been removed and replaced by its two children.

At any given stage of the algorithm, we have a current partition $\mathcal{P}$ consisting of intervals of two types: the collection $\mathcal{P}_G$ of *good intervals* I for which $e(u, I) \le \varepsilon$ and the collection $\mathcal{P}_B$ of *bad intervals* I for which $e(u, I) > \varepsilon$. The next step of the algorithm creates a new partition in which all the good intervals of $\mathcal{P}$ remain but each bad interval I is split and replaced by its two children. The process continues until $\mathcal{P}_B = \emptyset$. For the terminal partition $\mathcal{P}_\varepsilon$ and the resulting approximation $g_\varepsilon := \sum_{I \in \mathcal{P}_\varepsilon} m_I \chi_I$ (with again m_I the median of u on I), we clearly have

$$\|u - g_\varepsilon\|_{L^\infty(\Omega)} \le \varepsilon \tag{6.21}$$

and so our target accuracy has been met by the algorithm. Since u is continuous, the algorithm will terminate in a finite number of steps (dependent on u and ε). It has the advantage that the resulting algorithm can be described by a binary tree. Its nodes are the intervals that appear at any time in the refinement process. The leaves of the tree determine the intervals in the final partition $\mathcal{P}_\epsilon$. The tree structure is described graphically in Figure 6.4.

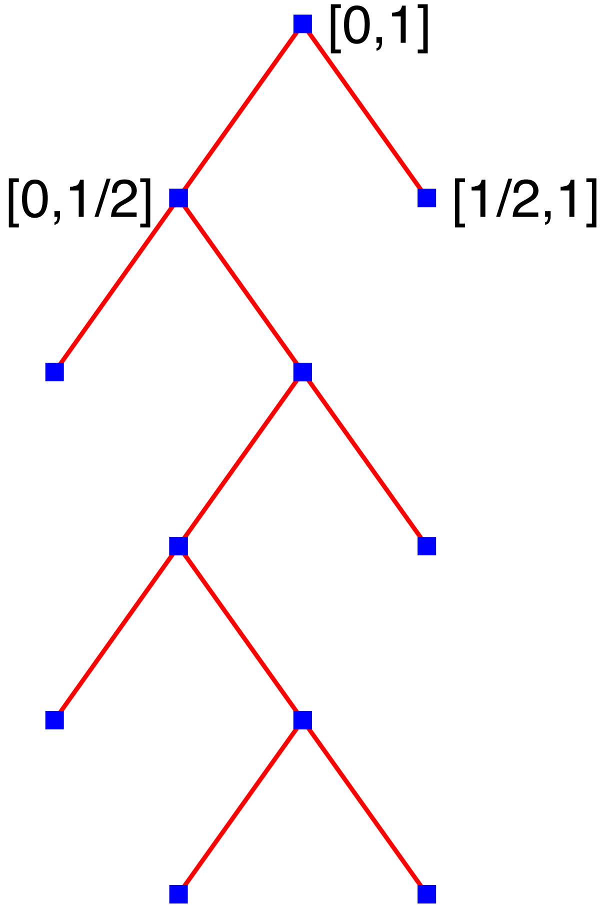

Figure 6.4: The figure shows a typical tree that arises in adaptive partitioning. The leaves of the tree give the final partition. In this case the intervals $[0, 1/4], [1/4, 5/16], [5/16, 11/32], [11/32, 3/8], [3/8, 1/2], [1/2, 1]$

The adaptive algorithm is more restrictive than the previously discussed nonlinear approximation by arbitrary piecewise constants functions, in the sense that the partitions which can arise only include *dyadic intervals*, i.e. intervals of the form $[2^{-j}k, 2^{-j}(k+1))$ with $j \in \mathbb{N}$ and $k \in \{0, \dots, 2^j - 1\}$.

The effectiveness of the above adaptive algorithm depends on how large the terminal partition $\mathcal{P}_\varepsilon$ is. We present one result which shows a typical way of bounding the cardinality of $\mathcal{P}_\varepsilon$. For this, we use the Hardy-Littlewood maximal function defined as follows. If $f \in L^1(\Omega)$ then

$$Mu(x) := \sup_{I \ni x} \frac{1}{|I|} \int_I |u(t)|\, dt, \quad x \in \Omega, \tag{6.22}$$

where the sup is taken over all intervals $I \subset \Omega$ that contain x. This maximal function plays an important role in harmonic analysis and its mapping properties are well-known. For example, when $1 < p \le \infty$,

$$Mu \in L^p(\Omega) \Leftrightarrow u \in L^p(\Omega), \tag{6.23}$$

and, when $p = 1$,

$$Mu \in L^1(\Omega) \Leftrightarrow |u| \log(1 + |u|) \in L^1(\Omega), \tag{6.24}$$

with equivalent norms.

We now show that whenever the maximal function of u' is in $L^1(\Omega)$ then

$$\#(\mathcal{P}_\varepsilon) \le \max\{1, \|Mu'\|_{L^1(\Omega)} \varepsilon^{-1}\}, \quad \epsilon > 0. \tag{6.25}$$

The proof of (6.25) is quite simple. We can assume $\#(\mathcal{P}_\varepsilon) > 1$. If $I \in \mathcal{P}_\varepsilon$ then, we consider its *parent* I', namely the interval which was split to generate I. This interval satisfies $e(u, I') > \varepsilon$ and therefore, for any $x \in I$,

$$|I| M(u')(x) = \frac{|I'|}{2} \frac{1}{|I'|} \int_{I'} |u'(t)| dt = \frac{V(u, I')}{2} \ge e(u, I') > \varepsilon. \tag{6.26}$$

Of course with a proper choice of $x \in I$, the left side does not exceed $\int_I M(u')(s) ds$. Hence summing the inequality $\varepsilon < \int_I M(u')(s) ds$ over all $I \in \mathcal{P}_\varepsilon$ and using the disjointness of the intervals in $\mathcal{P}_\varepsilon$, we arrive at (6.25).

If we choose $\varepsilon = \|M(u')\|_{L^1(\Omega)} n^{-1}$, we see that $\mathcal{P}_\varepsilon$ is a partition with at most n intervals which achieves the error $\|M(u')\|_{L^1(\Omega)} n^{-1}$. Thus the adaptive algorithm achieves an accuracy $\mathcal{O}(n^{-1})$ whenever $|u'| \log(1 + |u'|) \in L^1(\Omega)$ and in particular whenever $u' \in L^p(\Omega)$ for some $p > 1$. This is only a slightly stronger assumption than $u \in BV(\Omega)$ which was needed to guarantee the same rate of decay when allowing arbitrary partitions. On the other hand, it possible to check that the sole assumption $u \in BV(\Omega) \cap C(\Omega)$ is not sufficient to ensure that the adaptive algorithm achieves this rate.

The above results on piecewise constant approximation of univariate functions extend readily to higher order piecewise polynomials (see [23]). One can even stipulate global smoothness of the piecewise polynomial as long as this stipulation does not imply that the piecewise polynomial must reduce globally to a polynomial. We shall not formulate those results yet since they will be

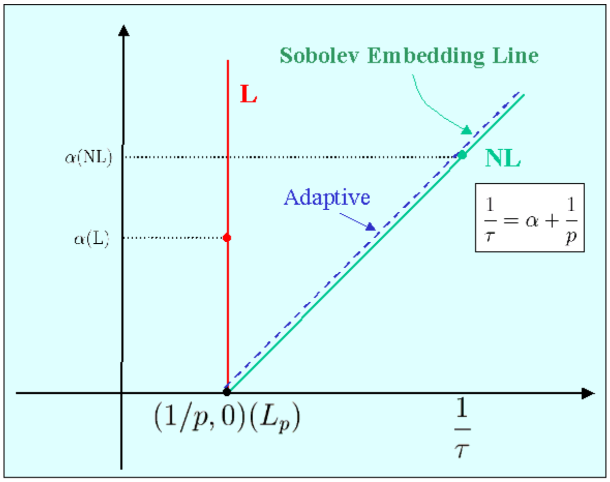

Figure 6.5: The figure depicts the smoothness conditions that guarantee a given approximation rate $O(n^{-\alpha})$ using adaptive partitioning. A sufficient condition that guarantees this order of approximation is that u is in a Besov class $B_q^\alpha(L_\tau)$ with $1/\tau < \alpha + 1/p$ as depicted by the dotted line. Note that this does not characterize the approximation spaces for adaptive approximation; these remain unknown.

a special case of multivariate approximation that we now discuss. Figure 6.5 gives a graphical description of the smoothness spaces that guarantee a given rate of approximation using adaptive approximation.

7 Multivarite approximation by piecewise polynomials

We next consider approximation of functions defined on a domain $\Omega \subset \mathbb{R}^d$. We shall restrict our discussion of multivariate approximation to the two settings most often used in numerical analysis: (i) piecewise polynomial approximation, (ii) n term approximation from a basis.

We begin with piecewise polynomials. The first issue to understand is what partitions can be allowed and still lead to a viable theory. The standard partitioning is to divide the given domain Ω into *cells* consisting of parallelpipeds or simplicies. Therefore, it is assumed initially that Ω is a polyhedral domain. The more generality allowed in the partitioning leads to weaker theory (in the sense of provable results) and more cumbersome numerical implementation.

A major restriction that is usually imposed on the cells in the allowable partitions is that they must be *uniformly shape preserving*. This means that there is an absolute constant c_0 such that each cell Q, that appears in any of the partitions, contains a sphere of radius $\geq c_0 \operatorname{diam}(Q)$. When this is applied to cells consisting of triangles or rectangles, this eliminates long thin triangles or rectangles.

We mention two common methods of partitioning in domains $\Omega \subset \mathbb{R}^d$.

Dyadic partitions: Let $\Omega = [0,1]^d$. For each $j \geq 0$, we define the set $\mathcal{D}_j$ of dyadic cubes in Ω of sidelength 2^{-j}. Thus, $\mathcal{D}_0 = \{\Omega\}$ and $\mathcal{D}_1$ is the set consisting of its 2^d children, and so on. These cubes are organized into a tree. Each dyadic cube Q has 2^d children and, as long as

$Q \neq \Omega$, it has one parent. Each set $\mathcal{D}_j$ is a (uniform) partition of Ω into 2^{jd} cubes. The set $\mathcal{D} := \cup_{j \geq 0}$ is the collection of all dyadic cubes in Ω.

We say a collection $\Gamma \subset \mathcal{D}$ is a tree if whenever $Q \in \Gamma$, its parent and all of its siblings are also in Γ. We obtain such trees by adaptive refinement. Starting with the collection of cubes consisting of Ω and its children, we mark some of its children for refinement and add the children of the marked cubes to the collection. Continuing in this way of marking and refining, we obtain the tree Γ. The set $\mathcal{L}(\Gamma)$ of *leaves* of Γ consists of those cubes in Γ whose children are not in Γ. The set of leaves of Γ form a partition of Ω.

Newest vertex bisection: This method for generating partitions is heavily used in adaptive finite elements. It is easiest to explain in the case $d = 2$. One begins with any polygonal domain Ω and an initial partitioning $\mathcal{P}_0$ of Ω into triangles. The sides of the triangles in $\mathcal{P}_0$ are given a labeling of 0 or 1 with the two stipulations: (i) the label depends only on the side and not the triangles (there may be two) containing the side, (ii) each side of a given triangle has exactly one side with label 0. The vertex opposite this side is called the *newest vertex.* It is known that it is always possible to make such an initial assignment [28],[6].

If a triangle T in $\mathcal{P}_0$ is to be subdivided then this is accomplished by bisecting the newest vertex or equivalently the side opposite it. This results in two triangles which are called the children of T. The new vertex created by this bisection is defined as the newest vertex for each of the two children that have been created. An adaptive partitioning is obtained by marking some of the triangles in $\mathcal{P}_0$ and then refining them. The partition $\mathcal{P}_1$ is then obtained from $\mathcal{P}_0$ by replacing each of the marked triangles by their two children. This procedure of marking and refining is repeated to generate the partitions $\mathcal{P}_2$, $\mathcal{P}_3$, and so on. Similar to the case of dyadic cubes, such partitioning can be associated to a dyadic forest (a dyadic binary tree emanating from each of the initial triangles in $\mathcal{P}_0$). If at each iteration, every triangle in $\mathcal{P}_j$ is refined to produce $\mathcal{P}_{j+1}$, then this is called *uniform partitioning.*

The *newest vertex bisection* procedure creates triangular partitions $\mathcal{P}_n$, $n = 0, 1, \ldots$, which may be non-conforming, i.e. they may have *hanging nodes.* They can be represented by a (finite) binary forest whose roots are the elements in $\mathcal{P}_0$ and whose leaves are the elements of $\mathcal{P}_n$. Each such forest is contained in an (infinite) master forest which consists of all triangles that may be generated from $\mathcal{P}_0$ by the newest vertex bisection procedure. We are mainly interested in *conforming* meshes (no hanging nodes), which correspond to a restricted class of finite binary forests. It was shown in [6] that any non-conforming partition $\mathcal{P}_n$ obtained by newest vertex bisection can be further refined so that the resulting partition is conforming and the cost of conforming refinement can be uniformly controlled. Namely, the cardinality of the conforming partition is bounded by a fixed multiple of the cardinality of $\mathcal{P}_n$. Newest vertex bisection can be extended to higher dimensions resulting in binary partitioning of simplicies. An important property of newest vertex bisection (in any space dimension d) is that the resulting simplicies are uniformly shape preserving.

Newest vertex subdivsion is used in many adaptive algorithms for the numerical solution of PDEs, see e.g. [29] for an up to date discussion of Adaptive Finite Element Methods (AFEMs) for elliptic problems. Recall that the main ingredient in adaptive algorithms is to identify the cells where the approximation error is the largest and mark them for refinement. The local approximation error on such a cell is assumed to be available in the approximation versions of the algorithm but is not in numerical PDE applications. So, the main challenge in numerical PDEs is to utilize the given computations to identify and mark the cells with large local error. This is accomplished through surrogates for the local error built from residuals. The fact that these surrogates are not equivalent to the true local error complicates the analysis of adaptive finite element methods.

7.1 Approximation classes for piecewise polynomials

Let us summarize the major results for linear and nonlinear approximation by piecewise polynomials.

7.1.1 Linear approximation

In linear approximation, we fix the sequence of partitions $\mathcal{P}_n$ and consider the space $\mathcal{S}^m(\mathcal{P}_n)$ of piecewise polynomials of order m (degree $m-1$) subordinate to these partitions. We assume, for the moment, that there is no stipulation of global smoothness on the functions in $\mathcal{S}^m(\mathcal{P}_n)$. We confine our attention to the case where the partitions $\mathcal{P}_n$ are obtained by uniform subdivision using either dyadic cube refinement or newest vertex refinement. Recall that in our discussion of approximation classes, we want spaces of dimension n and so for each n, we define $X_n := \mathcal{S}^m(\mathcal{P}_m)$ with m the largest integer such that $\dim(X_n) \leq n$. This means that these spline spaces are repeated when they occur in the sequence (X_n).

We fix the space $L_p(\Omega)$, $1 \leq p \leq \infty$, in which we measure distortion. In terms of direct theorems, the following is known. If $u \in W^s(L_p(\Omega))$ with the integer $s \leq m$, then the approximation error e_n satisfies

$$e_n(u) = e(u, X_n)_{L_p(\Omega)} \leq C(m,p)|u|_{W^s(L_p(\Omega))} n^{-s/d}, \quad n = 1, 2, \ldots. \tag{7.1}$$

More generally if u is in the Besov space $B^s_\infty(L_p(\Omega))$, where now s can be any number with $0 < s \leq m$, then

$$e_n(u) = e(u, X_n)_{L_p(\Omega)} \leq C(r,p)|u|_{B^s_\infty(L_p(\Omega))} n^{-s/d}, \quad n = 1, 2, \ldots. \tag{7.2}$$

Notice that (7.2) holds for all $0 < s \leq m$, whereas (7.1) is stated only for integers$m$.

The bound (7.2) does not characterize the approximation space $\mathcal{A}^{s/d}$, it only gives a sufficient condition for u to be in $\mathcal{A}^{s/d}$. It turns out that, we can actually characterize $\mathcal{A}^{s/d}$ for a certain range of s. Namely,

$$\mathcal{A}^{s/d} = B^s_\infty(L_p(\Omega)), \tag{7.3}$$

holds for a certain range of $0 < s < s^*$. The reason why we have to restrict the range of s is that we have assumed nothing about the smoothness of the piecewise polynomials. Obviously each piecewise polynomial subordinate to one of the partitions $\mathcal{P}_m$ can be approximated exactly once $n \geq m$. If we assume smoothness on the piecewise polynomials, we can increase the range of s for which (7.3) holds up the order of smoothness we impose. However, we then run the risk of destroying the local degrees of smoothness of the spline space. All these questions have a precise answer once the partitions and the continuity conditions on the piecewise polynomials is given.

7.2 Nonlinear piecewise polynomial approximation

We have seen in the univariate case the advantages of nonlinear approximation by piecewise polynomials. Namely, we have seen that we can obtain a specific rate of approximation with weaker assumptions on the target function u by allowing the partition for the piecewise polynomial be customized to the target function u. This paradigm persists in the case of multivariate approximation as we now briefly discuss.

The most commonly used method of nonlinear approximation for multivariate piecewise polynomials is adaptive approximation. This can be carried out for various settings in which the partitions are polyhedral cells. We restrict our discussion to simplicial decomposition and the newest vertex refinement strategy. We assume our domain Ω is a polyhedral domain which is initially partitioned into a finite set $\mathcal{P}_0$ of simplicial cells with an admissible labelling so that newest vertex subdivision can be applied.

If we execute an adaptive strategy of marking the cells with largest local error, we obtain adaptive partitions that can be associated to binary trees. We denote by $\mathfrak{P}_n$, the set of all partitions $\mathcal{P}$ that can be obtained from $\mathcal{P}_0$ by applying n newest vertex subdivision and further denote by $\Sigma_n = \Sigma_n^m$, the set of all piecewise polynomials of order m which are subordinate to one of the partitions of $\mathfrak{P}_n$. For the moment, we impose no smoothness conditions on the elements of Σ_n^m. Clearly, Σ_n is a nonlinear space.

The elements in Σ_n are generally obtained through adaptive paritioning, i.e. recursively marking certain cells for refinement. To describe the approximation power of this form of nonlinear approximation, we introduce the approximation error

$$\sigma_n(u)_{L_p(\Omega)} := \inf_{g\in\Sigma_n} \|u-g\|_{L_p(\Omega)}, \quad n \geq 1, \tag{7.4}$$

and the associated approximation classes $\mathcal{A}^r = \mathcal{A}^r((\Sigma_n), L_p(\Omega))$. The following general direct theorem holds

$$\sigma)n(u)_{L_p(\Omega)} \leq |u|_{B^s_\infty(L_\tau(\Omega))} n^{-s/d}, \quad n \geq 1, \tag{7.5}$$

provided $\tau > (s/d + 1/p)^{-1}$.

Let us make some comments on the above direct result. The condition that $u \in B^s_\infty(L_p(\Omega))$ with $\tau > (s/d + 1/p)^{-1}$, which guarantees the approximation rate $O(n^{-s/d})$ corresponds to requiring that this Besov space lies above the Sobolev embedding line. Thus, the analogue of Figure 6.3 holds with the following modifications. The Sobolev embedding line is now $1/\tau = s/d + 1/p$. The reason we require that the Besov space lies strictly above the embedding line is because we are using a form of adaptive approximation.

There is no exact characterization of the approximation classes $\mathcal{A}^r$ in this setting. However, when global smoothness is imposed on the approximants in Σ_n, then there is a range (depending on the smoothness imposed) for which one almost has a converse. Namely, if u is in $\mathcal{A}^{s/d}$ then it must be in the Besov space $B^s_\tau(L_\tau(\Omega))$ with $1/\tau = s/d + 1/p$. For a proof of the above results and further elaboration the reader should consult [7].

7.3 Implications of pp theory

Approximation theory has played a significant role in proving the optimality of adaptive finite element algorithms. It is shown in [6, 17] that for a class of elliptic problems on a Lipschitz domain, certain AFEMs, built on newest vertex refinements, provide an approximation error $O(n^{-r})$ whenever the solution u is in the approximation class $\mathcal{A}^r$ for adaptive approximation. In this sense, these AFEMs cannot be improved. That is, once the decision was made to use piecewise polynomial approximation built on newest vertex refinement, no algorithm can do better in terms of rate of approximation. The results of [6] have been extended and improved in many directions (see [29]).

Beyond proving optimality, approximation theory gives an a priori determination of the class of problems for which adaptive algorithms perform well. This is accomplished by first showing [7], that the approximation classes $\mathcal{A}^r$ for adaptive approximation are related to Besov smoothness. Namely, a function u is in $\mathcal{A}^r$ whenever it is in a Besov space $B^s_q(L_p(\Omega))$ which compactly embeds into $H^1(\Omega)$. By the Sobolev embedding theorem such a compact embedding holds if and only if $s - 1 > d/p - d/2$. Then, regularity theorems for elliptic problems (see [18]) clarify for which elliptic problems, i.e. for which right sides f and which diffusion coefficients, the solution u lies in such Besov spaces. We will discuss elliptic problems in more detail in our last lecture.

8 Basis approximation

Another popular class of approximation methods begins with a basis of functions for X and uses linear combinations of these basis elements for the approximation. To understand the distinction between linear and nonlinear methods in this setting, we consider the simplest case where $X = \mathcal{H}$ is a separable Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|u\|_{\mathcal{H}} := \langle u, u\rangle^{1/2}$. The prototypical example for $\mathcal{H}$ would be the space $L_2(\Omega)$. Let $\Psi = \{\psi_1, \psi_2, \ldots\}$ be an orthonormal basis for $\mathcal{H}$ which means that each element $u \in \mathcal{H}$ has a unique expansion

$$u = \sum_{k=1}^{\infty} c_k \psi_k, \quad c_k := c_k(u) := \langle u, \psi_k\rangle. \tag{8.1}$$

For an important concrete example, the reader may keep in mind the space $\mathcal{H} = L_2[-\pi, \pi]$ of square integrable 2π periodic functions and the Fourier basis.

8.1 Linear basis approximation

In linear approximation, we consider the nested linear spaces

$$\mathcal{H}_n := \operatorname{span}\{\psi_1, \ldots, \psi_n\}, \quad n \geq 1. \tag{8.2}$$

Given $u \in \mathcal{H}$, the function $P_n u := \sum_{k=1}^n c_k(u)\psi_k$ is the best approximation to f from $\mathcal{H}_n$. The operator P_n is the orthogonal projector onto $\mathcal{H}_n$.

The error we incur in this form of linear approximation is given by

$$e_n(f) := e_n(u)_{\mathcal{H}} := \|u - P_n u\|_{\mathcal{H}} = \left(\sum_{k=n+1}^{\infty} |c_k(u)|^2\right)^{1/2}. \tag{8.3}$$

We notice that $e_n(u)$ tends monotonically to 0. We are in the wonderful situation of having an explicit formula for the error of approximation in terms of the coefficients $c_k(f)$. From this, one can easily deduce that for any $r > 0$, the approximation space $\mathcal{A}^r = \mathcal{A}^r(\mathcal{H}, (\mathcal{H}_n)_{n\geq 1})$ is characterized by the condition

$$\sum_{2^j \leq k < 2^{j+1}} |c_k(u)|^2 \leq M^2 2^{-2jr}, \quad j \geq 0, \tag{8.4}$$

and the smallest constant $M = M(f)$ in (8.4) is equivalent to the $\mathcal{A}^r$ norm. The condition (8.4) reflects a decay property of the coefficients $c_k(f)$ in an averaged sense. One can view this as a smoothness condition whose exact form depends on the basis Ψ. In the case of the Fourier basis, note that a slightly stronger condition than (8.4) is that the series $\sum_{k\geq 1} |k^r c_k(u)|^2$ converges. When r is an integer this means that all derivatives $u^{(m)}$ for $0 \leq m \leq r$ belong to $L^2[-\pi, \pi]$, or equivalently u belongs to the periodic Sobolev space $W^r(L_2[-\pi, \pi])$. It is also possible to prove that the condition (8.4) is equivalent to a slightly weaker smoothness condition, namely that u belongs to the Besov space $B^r_\infty(L^2[-\pi, \pi])$.

8.2 Best n-term approximation

In nonlinear approximation, rather than approximate f by the first n terms of its expansion (8.1), we consider approximation by any n terms. Namely, we define the set

$$\Sigma_n := \{\sum_{j\in\Lambda} d_j \psi_j \ : \ \#(\Lambda) \leq n\}, \tag{8.5}$$

of all n-term linear combinations of elements of Ψ. Notice that the space Σ_n is not linear. If we add two elements from Σ_n, we will generally need $2n$ terms to represent the sum. Another view of n-term approximation is that we approximate any function f by the elements of a linear space W_n spanned by n basis functions, however this space can be chosen depending on f, that is, it is not fixed in advance as was the case for linear approximation.

It is very easy to describe the best approximation to f from Σ_n and the resulting error of approximation. Given any sequence $(a_j)_{j\geq 1}$ of real numbers which tend to zero as $j \to \infty$, we denote by $(a_k^*)_{j\geq 1}$ the decreasing rearrangement of the $|a_j|$. Thus, a_k^* is the k-th largest of these numbers. For each k, we can find a λ_k such that $a_k^* = |a_{\lambda_k}|$ but the mapping $k \mapsto \lambda_k$ is not unique because of possible ties in the size of the entries. The following discussion is impervious to such differences. If we apply rearrangements to the coordinates $\{c_j(u)\}_{j\geq 1}$ and denote by $\Lambda_n := \Lambda_n(u) := \{j_1, \ldots, j_n\}$ the indices of a set of n-largest coefficients, then the best approximation to $u \in \mathcal{H}$ from Σ_n is given by the function

$$G_n u := \sum_{k=1}^{n} c_{j_k}(u)\psi_{j_k} = \sum_{j\in\Lambda_n} c_j(u)\psi_j, \tag{8.6}$$

and the resulting error of approximation is

$$\sigma_n(u)^2 = \sigma_n(u)_{\mathcal{H}}^2 = \|u - G_n u\|^2 = \sum_{j\notin\Lambda_n} |c_j(u)|^2 = \sum_{k>n} (c_k^*(u))^2. \tag{8.7}$$

While best approximation from Σ_n is not unique, the approximation error $\sigma_n(u)_{\mathcal{H}}$ is uniquely defined. Also, note that $\sigma_n(u)_{\mathcal{H}} = \sigma_n(\tilde{u})_{\mathcal{H}}$ if u and $\tilde{u}$ have the same coefficients up to a permutation of the indices.

We can use (8.7) to characterize $\mathcal{A}^r(\mathcal{H},(\Sigma_n)_{n\geq 1})$ for any $r > 0$ in terms of the coefficients $c_j(u)$. Given such an $r > 0$, we define p by the formula

$$\frac{1}{p} = r + \frac{1}{2}. \tag{8.8}$$

Notice that $p < 2$. The space $w\ell_p$ (*weak* ℓ_p) is defined as the set of all $\mathbf{a} = (a_j)_{j\geq 1}$ whose decreasing rearrangement $(a_k^*)_{k\geq 1}$satisfies

$$k^{1/p} a_k^* \leq M, \quad k \geq 1. \tag{8.9}$$

and the smallest $M = M(\mathbf{a})$ for which (8.9) is valid is the quasi-norm $\|\mathbf{a}\|_{w\ell_p}$ of $\mathbf{a}$ in this space. Notice that $w\ell_p$ contains ℓ_p and is slightly larger since it contains sequences whose rearrangement behaves like $k^{-1/p}$ which barely miss being in ℓ_p. We claim that, with $\mathbf{c} := \mathbf{c}(u) := \{c_j(u)\}_{\jmath\geq 1}$,

$$\mathcal{A}^r := \mathcal{A}^r(\mathcal{H},(\Sigma_n)_{n\geq 1}) = \{u \ : \ \mathbf{c}(u) \in w\ell_p\}, \tag{8.10}$$

and $\|\mathbf{c}(u)\|_{w\ell_p}$ is equivalent to $\|u\|_{\mathcal{A}^r}$.

Indeed, if $\mathbf{c}(u) \in w\ell_p$, then for any $n \geq 1$, we have

$$\sigma_n(u) = \sum_{k>n} (c_k^*(u))^2 \leq \|\mathbf{c}(f)\|_{w\ell_p}^2 \sum_{k>n} k^{-2r-1} \leq \frac{1}{2r}\|\mathbf{c}(u)\|_{w\ell_p}^2 n^{-2r}. \tag{8.11}$$

In addition,

$$\|u\|_{\mathcal{H}}^2 = \|\mathbf{c}(u)\|_{\ell^2}^2 \leq \|\mathbf{c}(u)\|_{w\ell^p}^2 \sum_{k\geq 1} k^{-2r-1} \leq (1 + \frac{1}{2r})\|\mathbf{c}(u)\|_{w\ell_p}^2. \tag{8.12}$$

This shows that $\|u\|_{\mathcal{A}^r} \leq (1 + \frac{1}{2r})^{1/2}\|\mathbf{c}(u)\|_{w\ell_p}$.

To reverse this inequality, we note that for any $k \geq 1$, the monotonicity of $\mathbf{c}^*(u)$ gives

$$2^j (c_{2^{j+1}}^*(u))^2 \leq \sum_{k=2^j+1}^{2^{j+1}} (c_k(u)^*)^2 \leq \sigma_{2^j}(u)^2 \leq |u|_{\mathcal{A}^r}^2 2^{-2jr}. \tag{8.13}$$

For any n, we choose j so that $2^j \leq n < 2^{j+1}$. If $j > 0$, we obtain from the monotonicity of $\mathbf{c}^*(u)$ that

$$c_n^*(u) \leq c_{2^j}^*(u) \leq 2^{r+1/2}|u|_{\mathcal{A}^r} 2^{-(r+1/2)j} = 2^{1/p}|u|_{\mathcal{A}^r} 2^{-j/p} \leq 2^{2/p}|u|_{\mathcal{A}^r} n^{-1/p}. \tag{8.14}$$

On the other hand, we clearly have

$$c_1^*(u) \le \|u\|_{\mathcal{H}} \le \|u\|_{\mathcal{A}^r}. \tag{8.15}$$

This gives $\|\mathbf{c}(u)\|_{w\ell_p} \le 2^{2/p}\|u\|_{\mathcal{A}^r}$ and completes the proof of the equivalence.

Let us conclude this example by making some remarks on what we have shown. As we know, the condition $\mathbf{c}(u) \in w\ell_p$, $1/p = r + 1/2$, is weaker than (8.4). In particular, it is independent of the ordering of the index set and hence for a fixed ordering, it does not imply any form of decay of the coefficient sequence.

The distinction between linear and nonlinear approximation manifests itself more clearly when considering concrete examples. For example, in the case of Fourier series discussed earlier, the smoothness conditions that ensure $\mathbf{c}(u) \in w\ell_p$ are typically weaker than those ensuring (8.4).

The particular case where Ψ is a wavelet basis and $X = L_p$ is very popular in image compression and statistics. Keeping with our theme of piecewise constant univariate approximation, let us mention the simplest wavelet. The Haar wavelet is the function

$$H(x) := \begin{cases} -1, & 0 \le x < 1/2 \\ +1, & 1/2 \le x \le 1, \end{cases} \tag{8.16}$$

Let $\mathcal{D}$ denote the set of all dyadic intervals in $\Omega = [0,1]$. For each $I \in \mathcal{D}$, $I = 2^{-k}[j, j+1]$, we can define H_I as the scaled version

$$H_I(x) := 2^{k/2}\psi(2^k x - j) \tag{8.17}$$

of H fit to the interval I. It is easy to see that $\Psi := \{\varphi\} \cup \{H_I\}_{I\in\mathcal{D}}$ is an orthonormal basis for $L_2[0,1]$. Indeed, by our previous discussion, the linear combinations of these functions are dense in $L_2[0,1]$ and it is easy to check that they are orthonormal. Notice, that it was more convenient to index the Haar basis by the dyadic intervals $I \in \mathcal{D}$ rather than the integers $j \ge 1$.

Every function $f \in L_2[0,1]$ has an expansion

$$f = c_0\chi_{[0,1]} + \sum_{I\in\mathcal{D}} c_I\psi_I. \tag{8.18}$$

If we consider the linear basis approximation using this basis, it is essentially the same as using approximation using equally spaced partitions as was done in §6.1. In fact the space spanned by $\chi_{[0,1]}$ and the H_I with $|I| < 2^{-k}$ is the same as $\mathcal{S}_{2^k}$.

The space Σ_n for the basis problem is similar to that for nonlinear piecewise constant approximation. Again, one confronts a difficulty in implementing the nonlinear basis approximation since one cannot search over all coefficients of f in order to find its decreasing rearrangement. The remedy for this is to restrict the sets Λ in the nonlinear basis problem to form a tree of dyadic intervals. The resulting approximation tool is called *tree approximation.* Tree approximation, using the Haar basis, is almost identical to adaptive piecewise constant approximation. One should note that when measuring the distortion in $L_2[0,1]$, tree approximation has a very simple and numerically friendly implementation through thresholding. For any $\epsilon > 0$, we take the smallest tree that contains all Haar coefficients c_I with $|c_I| > \epsilon$. The approximation class for this implementation of tree approximation is then analogous to the approximation class for adaptive piecewise constant approximation [15].

The Haar wavelet, despite its simplicity, is seldom used in numerical settings because of its lack of smoothness. This leads to the question of whether H can be replaced by a smoother function ψ such that the functions ψ_I obtained by replacing H by ψ in (8.17) form an orthonormal family. The answer was provided by Ingrid Daubechies (see [19]) who constructed a family of such function ψ each of which provides the desired orthogonality. Any order of smoothness can be required of ψ but at the expense of increasing the support of ψ.

The Daubechies wavelets and their multivariate analogues are staples in signal processing and statistical estimation. Imaging tasks, such as compression [21, 15] and denoising [26] can be implemented numerically by application of finite filters and simple coefficient operations such as thresholding and quantization. Wavelets are also used for solving operator equations [12, 13, 14]. The theory of nonlinear basis approximation, especially tree approximation, can be used to show that the resulting algorithms in each of these disciplines are order optimal. Approximation theory also describes the properties of the target function which give a prescribed distortion rate as members of certain Besov spaces.

9 The Curse of Dimensionality

Having given this very brief and coarse review of classical approximation theory, let us now turn our focus to the main point of emphasis of these lectures. We have seen that the performance of approximation methods is closely tied to smoothness. Regardless of whether we use linear or nonlinear methods of approximation, the rate of approximation we obtained is governed by the membership of the target function in a classical smoothness space. For example, when using classical linear methods of approximation to capture a function in the sense of $L_p(\Omega)$, Ω a domain in $\mathbb{R}^d$, by piecewise polynomials or other numerically realizable methods such as wavelet expansions, then whenever u is in the Besov space $B^s_\infty(L_p(\Omega))$ (or in $W^s(L_p(\Omega))$, when s is an integer), it can be approximated with error

$$E_n(u)_{L_p(\Omega)} \le C(s,d)|u|_{B^s_\infty(L_p(\Omega))} n^{-s/d}. \tag{9.1}$$

We even know that this is a necessary condition in many settings. This result points out the debilitating effect of the space dimenson d. If d is large, the requirement on s must increase to guarantee a given rate. For example, to attain the asymptotic rate $O(n^{-1})$ requires that u have d derivatives in $L_p(\Omega)$. This does not even take into consideration the effect of the constant $C(s,d)$ appearing in (9.1). It is known that this constant also increases with d.

The situation for nonlinear methods is slightly better in that we can achieve the approximation rate $O(n^{-s/d})$ with a slightly weaker Besov space condition such as $u \in B^s_\infty(L_\tau(\Omega))$ provided $\tau > (s/d + 1/p)^{-1}$. Since τ can be considerably smaller than p, this is a weaker requirement on u. Note, however, that this improvement given by nonlinear methods does not avoid the curse. We still need an inordinately high amount of smoothness to achieve reasonable approximation error decay. It is just that the form of this smoothness is significantly weaker.

9.1 Widths and Entropy of Classes

The astute reader can rightfully question the above discussion by saying it only applies to the classical methods of approximation like piecewise polynomials or wavelets. Maybe there is some other miraculous way of approximating that avoids this curse. To dispell all doubt on this matter, we turn now to the subject of widths and entropy which examines the optimal performance of all methods of approximation and thereby will show that no methods can avoid this curse.

9.2 Kolmogorov entropy

Let us first consider entropy. Suppose that K is a compact set in the Banach space X with norm $\|\cdot\|_X$. If $\epsilon > 0$, we consider all possible coverings of $K \subset \bigcup_{i=1}^m B(x_i, \epsilon)$ using balls $B(x_i, \epsilon)$ of radius ϵ with centers $x_i \in X$. The smallest number $m = N_\epsilon(K)_X$, for which such a covering exists, is called the covering number of K. The Kolmogorov entropy of K is then defined as

$$H_\epsilon(K)_X := \log_2(N_\epsilon(K))_X. \tag{9.2}$$

The Kolmogorov entropy measures the size or massivity of K. It has another important property of determining optimal encoding of the elements of K. Namely, if $x \in K$ then we can assign to x the binary bits of an index i for which $x \in B(x_i, \epsilon)$. Each x is then encoded to accuracy ϵ with $\leq \lceil H_\epsilon(K)_X \rceil$ bits and no other encoder can do better for K.

It is frequently more convenient to consider the entropy numbers

$$\epsilon_n(K)_X := \inf\{\epsilon : H_\epsilon(K)_X \leq n\}. \tag{9.3}$$

Typically, $\epsilon_n(K)_X$ decay like n^{-r} for standard compact sets. Not only does $\epsilon_n(K)_X$ tell us the minimal distortion we can achieve with n bit encoding, it also says that any numerical algorithm which computes an approximation to each of the elements of K to accuracy $\epsilon_n(K)_X$ will require at least n operations.

An important issue for us is what are the entropy numbers of the classical smoothness spaces. if K is the unit ball $U(W^s(L_p(\Omega))$ of a Sobolev space, or a unit ball $U(B_q^s(L_p(\Omega))$ of a Besov space, then for any Lebesgue space $X = L_\mu(\Omega)$,

$$\epsilon_n(K)_X \geq Cn^{-s/d}, \quad n = 0, 1, \ldots. \tag{9.4}$$

This result manifests the massivity of these compact sets as the dimension d increases.

9.3 Widths

There are several types of widths. The most prominent of these is the Kolmogorov width which measures how well K can be approximated through linear spaces of fixed dimension n. Given the Banach space X in which we wish to measure error and the compact set $K \subset X$, it is defined by

$$d_n(K)_X := \inf_{\dim(Y)=n} \sup_{u\in K} \inf_{g\in Y} \|u - g\|_X, \quad n = 0, 1, \ldots. \tag{9.5}$$

In other words, given K and any n dimensional linear space, we look at how well Y approximates the elements in K by measuring the worst performance on K. Then we optimize over the best

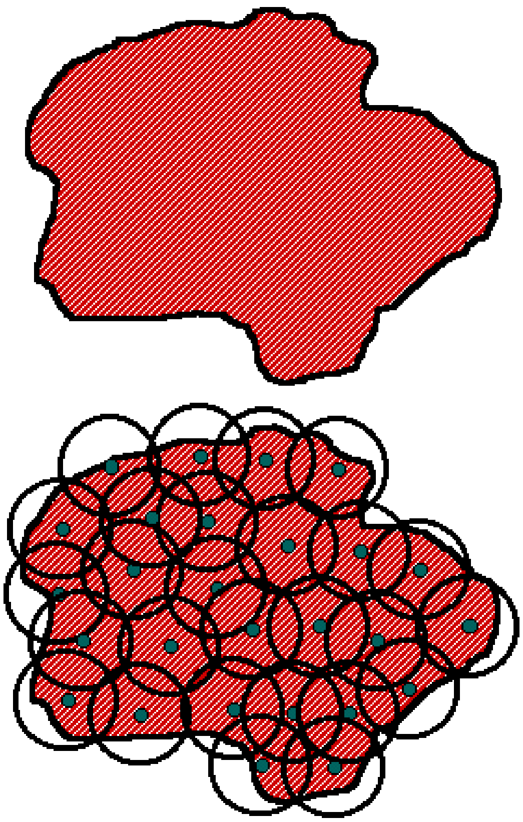

Figure 9.6: A compact set K and its ϵ cover

Y. What is relevant for us is that if K is the unit ball $U(W^s(L_p(\Omega))$ of a Sobolev space, or a unit ball $U(B^s_q(L_p(\Omega))$ of a Besov space, then for any Lebesgue space $X = L_\mu(\Omega)$,

$$d_n(K)_X \geq C n^{-s/d}, \quad n = 0, 1, \dots . \tag{9.6}$$

So the curse cannot be avoided by some judicious choice of n dimensional spaces.

9.3.1 Nonlinear widths

There have been several definitions of nonlinear widths that have been proposed to measure optimal performance of nonlinear methods. However, for us, the following definition of manifold width [20] will be the most useful since it matches well numerical algorithms based on nonlinear approximation. To define this width, we consider two continuous functions. The first function b maps each element $x \in K$ into $\mathbb{R}^n$ and the second function M maps $\mathbb{R}^n$ into the set $\mathcal{M}$ (which we view as an n-dimensional manifold although we make no assumptions about the smoothness of the image $\mathcal{M}$). The manifold width of the compact set K is then defined by

$$\delta_n(K)_X := \inf_{M,b} \sup_{x \in K} \|x - M(b(x))\|_X. \tag{9.7}$$

For typical compact sets K of functions, the manifold widths behave like the entropy numbers. For example, if K is the unit ball of any Besov or Sobolev space of smoothness s which compactly embeds into $L_p(\Omega)$ with $\Omega \subset \mathbb{R}^N$, then (see [22])

$$C_0 n^{-s/N} \leq \delta_n(K)_{L_p(\Omega)} \leq C_1 n^{-s/N}. \tag{9.8}$$

We see in (9.8) the curse of dimensionality. In order to obtain just moderate rates of convergence with $n \to \infty$ we need s to be comparable with N.

9.4 Model classes for functions in high dimension

We have seen that classical smoothness spaces suffer from 'the curse of dimensionality'. Does this mean that there is no hope to numerically solve problems in high dimensions? This would be the case if all we know about the function u is that it has classical smoothness. On the other hand, we expect that real world problems can be resolved numerically and so the solutions to such problems must have some other properties that make them easier to capture numerically. This train of thought has led to the creation of many new model classes to describe functions in high dimension. These center around notions such as sparsity, variable reduction, or reduced modelling. These topics will be the focal point of the lectures that follow.

References

[1] R. Adams, Sobolev Spaces, Academic Press, New York, 1975,

[2] C. de Boor, A Practical Guide to Splines, Applied Mathematical Sciences, Vol. 27, Springer Verlag, Berlin, 1978.

[3] G. Baker and P. Graves-Morris, Padé Approximants, Cambridge Univ. Press, 1996.

[4] C. Bennett and R. Sharpley, Interpolation of Operators, Academic Press, 1988.

[5] J. Bergh and J. Löfström, Interpolation Spaces, Grundlehren, vol. 223, Springer Verlag, Berlin, 1976.

[6] Peter Binev, W. Dahmen, and R. DeVore, *Adaptive Finite Element Methods with Convergence Rates*, Numerische Mathematik, **97**(2004) 219–268.

[7] Peter Binev, W. Dahmen, R. DeVore, and P. Petrushev, *Approximation Classes for Adaptive Methods*, Serdica Math. J., **28**(2002), 391–416.

[8] D. Braess, Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics, Cambridge University Press, 2007.

[9] S. Brenner and R. Scott, The Mathematical Theory of Finite Element Methods, Texts in Applied Mathematics, Vol. 15, Springer Verlag, Berlin.

[10] P. Ciarlet, The Finite Element Method for Elliptic Problems, Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1978.

[11] A. Cohen, Numerical analysis of wavelet methods, Studies in Mathematics and its Applications, Elsevier, Amsterdam, 2003.

[12] A. Cohen, W. Dahmen and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: convergence rates* , Math. Comp., **70** (2000) 27–75.

[13] A. Cohen, W. Dahmen and R. DeVore , *Adaptive wavelet methods II - Beyond the elliptic case*, J. FoCM, **2**(2002), 203–245.

[14] A. Cohen, W. Dahmen and R. DeVore , *Adaptive Wavelet Schemes for Nonlinear Variational Problems*, SIAM J. Numer. Anal., **41**(2003), 1785–1823.

[15] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore, *Tree Approximation and Encoding*, ACHA, **11**(2001) 192–226.

[16] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard, *Maximal Spaces with given rate of convergence for thresholding algorithms*, ACHA, **11**(2001) 167–191.

[17] A. Cohen, R. DeVore, and R. Nochetto, *Convergence Rates of AFEM with H^{-1} Data*, J. FoCM **12**(2012), 671–718.

[18] S. Dahlke and R. DeVore, *Besov regularity for 2-D elliptic boundary value problems with variable coefficients*, Communication in PDEs, **22**(1997) 1–16.

[19] I. Daubechies, Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61, 1992.

[20] R. DeVore, R. Howard and C. Micchelli, *Optimal non-linear approximation*, Manuscripta Math., **63** (1989) 469–478. 59.

[21] R. DeVore, B. Jawerth and B. Lucier, *Image compression through transform coding*, IEEE Proceedings on Information Theory, **38**(1992), 719–746.

[22] 78. R. DeVore, G. Kyriazis, D. Leviatan, and V.M. Tikhomirov, *Wavelet compression and nonlinear n-widths*, Advances in Computational Math., **1** (1993) 197–214.

[23] R. DeVore and G.G. Lorentz, Constructive Approximation,Grundlehren, vol. 303, Springer Verlag, New York, 1996.

[24] R. DeVore and B. Lucier, *Wavelets*, Acta Numerica, Volume 1 (1991), 1–56.

[25] R. DeVore and V. Popov, *Interpolation spaces and nonlinear approximation*, in Function Spces and Approximation, M. Cwikel et al. (eds.), Springer Lecture Notes in Mathematics, vol. 1302, Springer, Berlin, 1988, 191–205.

[26] D. Donoho and I. Johnstone, *Ideal Spatial Adaptation by Wavelet Shrinkage*, Biometrika, **81** (1994), 425–455.

[27] J. P. Kahane, Teoria Constructiva de Functiones, Course Notes, University de Buenos Aires, 1961.

[28] W.F. Mitchell, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math Softw., 15 (1989), 326–347.

[29] R, Nochetto, K. Siebert and A. Veeser, *Theory of adaptive finite element methods: an introduction*, in Multiscale, Nonlinear and Adaptive Approximation, R. DeVore and A. Kunoth (eds.), Springer, Berlin, 2009, 409–542.

[30] P. Petrushev, *Direcct and inverse theorems for spline and rational approximation and Besov spaces*, in Function Spces and Approximation, M. Cwikel et al. (eds.), Springer Lecture Notes in Mathematics, vol. 1302, Springer, Berlin, 1988, 363–377.

[31] I. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions*, Quart. Appl. Math., **4**(1946), 45-99 and 112-141.

[32] L. Schumaker, Spline Functions: Basic Theory Wiley, 1980.

[33] V. Temlyakov, Greedy Approximation, Cambridge University Press, 2011.

[34] H. Wendland, Scattered Data Approximation, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2005.