Appendix E

Polynomials, Splines and Wavelets

This appendix gathers results on three classes of functions used to approximate other functions: Bernstein polynomials, splines and wavelets. We focus on properties that are used throughout the main text.

E.1 Polynomials

Polynomials are widely used because of their simplicity and approximation ability. The classical Weierstrass theorem asserts that every continuous function on a compact interval in the real line can be uniformly approximated by polynomials. To approximate a general function, the degree of the polynomials must tend to infinity, where the rate of approximation as a function of the degree may be arbitrarily slow. However, the uniform distance between a function f in the Hölder space $\mathfrak{C}^\alpha[0,1]$ and the closest polynomial of degree k is of the order $k^{-\alpha}$.

Proposition E.1 (Polynomial approximation) *There exists a constant D that depends only on α such that for every $f \in \mathfrak{C}^\alpha[0,1]$ and $k \in \mathbb{N}$ there exists a polynomial P of degree k such that $\|f - P\|_\infty \le Dk^{-\alpha}\|f\|_{\mathfrak{C}^\alpha}$.*

A related result is that a periodic function in $\mathfrak{C}^\alpha[0,1]$ can be approximated with the same order of accuracy $k^{-\alpha}$ by trigonometric polynomials (Jackson 1912; these are polynomials in sine and cosine functions; a Fourier series is an example of such a polynomial, but gives the approximation only up to logarithmic factor). The relation between the two settings arises by considering ordinary polynomials in $z = e^{it}$, which are trigonometric polynomials as functions of t.

The remainder of this section is concerned with Bernstein polynomials, which are a special type of polynomial that can be obtained in a constructive manner from a given function. Even though their accuracy of approximation is suboptimal, they are interesting for retaining properties such as monotonicity or positiveness.

For a continuous function $F\colon (0,1] \to \mathbb{R}$, the associated *Bernstein polynomial* is defined as

$$B(x; k, F) = \sum_{j=0}^{k} F\left(\frac{j}{k}\right)\binom{k}{j} x^j (1-x)^{k-j}. \tag{E.1}$$

As $k \to \infty$ the functions $x \mapsto B(x; k, F)$ converge uniformly to F. This can be derived from the law of large numbers by the representation

$$B(x;k,F) = \mathrm{E}\Big[F\Big(\frac{J}{k}\Big)\Big], \qquad J \sim \mathrm{Bin}(k,x).$$

The result follows since $J/k \to x$ in probability, uniformly in x. The representation also shows that if $F(x)$ takes values in an interval $[c_1, c_2]$, so does $B(x;k,F)$.

The derivative of $B(x;k,F)$ in $(0,1)$ is

$$b(x;k,F) = \sum_{j=1}^{k}\Big(F\Big(\frac{j}{k}\Big) - F\Big(\frac{j-1}{k}\Big)\Big)\,\mathrm{be}(x;j,k-j+1), \tag{E.2}$$

where $\mathrm{be}(x;a,b)$ stands for the beta density $\mathrm{be}(x;a,b) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))\,x^{a-1}(1-x)^{b-1}$. In particular, the derivative is positive if F is monotone. Further, if F is a cumulative distribution function on $[0,1]$, then $B(1;k,F) = 1$, so that $B(x;k,F)$ is also a cumulative distribution function on $[0,1]$. These two shape-preserving properties of Bernstein polynomials are attractive for the purpose of approximating a function, as natural constraints can be maintained. Also observe that if F is a probability distribution on $[0,1]$ with $F(0) = 0$, then $B(x;k,F)$ has a density representable as a mixture of beta densities $\mathrm{be}(x;j,k-j+1)$, $j = 1,\ldots,k$.

More generally, given a weight sequence $w = (w_j\colon 1 \le j \le k)$, $k \in \mathbb{N}$, we define a *Bernstein polynomial density corresponding to* w by

$$b(x;k,w) := \sum_{j=1}^{k} w_j\,\mathrm{be}(x;j,k-j+1) = k\sum_{j=1}^{k} w_j\binom{k-1}{j-1}x^{j-1}(1-x)^{k-j}, \tag{E.3}$$

If w is a probability vector, then p is a probability density function. Note the abuse of the notation $b(\cdot;k,\cdot)$ applied to both a function and a weight sequence in its third argument.

Lemma E.2 *For every $w \in \mathbb{R}^k$, $\|b(\cdot;k,w)\|_\infty \le k\|w\|_\infty$ and*

$$\|b(\cdot;k,w) - b(\cdot;k,w')\|_1 \le \|w - w'\|_1.$$

Proof In view of (E.3) the norm $\|b(\cdot;k,w)\|_\infty$ is bounded by

$$k\,\max_j |w_j|\,\sup_x \sum_{i=0}^{k-1}\binom{k-1}{i}x^i(1-x)^{k-1-i}.$$

This is equal to $k\|w\|_\infty$ by the binomial formula. The second relation follows by integrating the first middle expression in (E.3) after taking absolute difference at w and w'. $\square$

The following result describes the accuracy of approximation of Bernstein polynomials at α-smooth densities in the sense of Definition C.4. Recall that the Lipschitz constant of order $\alpha \in (0,1]$ of a continuous function f is given by $L_\alpha(f) = \sup\{|f(x_1) - f(x_2)|/|x_1 - x_2|^\alpha\colon x_1 \ne x_2\}$. In the lemmas below the target function f need not be a probability density.

Lemma E.3 *For $f \in \mathfrak{C}^\alpha[0,1]$ and $\alpha \in (0,2]$, define $M_\alpha(f)$ to be $L_\alpha(f)$ if $\alpha \le 1$ and to be $L_{\alpha-1}(f') + \|f'\|_\infty$ if $1 < \alpha \le 2$. Then $\|f - b(\cdot;k,F)\|_\infty \le 3M_\alpha(f)k^{-\alpha/2}$ for any $\alpha \in (0,2]$, where F is a primitive function of f.*

Proof By (E.3), with J a variable with the $\mathrm{Bin}(k-1,x)$-distribution,

$$b(x;k,F)=k\mathrm{E}\Big[F\Big(\frac{J+1}{k}\Big)-F\Big(\frac{J}{k}\Big)\Big]. \tag{E.4}$$

First consider the case that $\alpha\in(0,1]$. By the mean value theorem $F((y+1)/k)-F(y/k)=k^{-1}f(\xi_y)$, for some $\xi_y\in[y/k,(y+1)/k]$. Consequently, by (E.4),

$$\begin{aligned}|f(x)-b(x;k,F)|&=|f(x)-\mathrm{E}f(\xi_J)|\le\mathrm{E}|f(x)-f(\xi_J)|\\&\le L_\alpha(f)\mathrm{E}|x-\xi_J|^\alpha\le L_\alpha(f)(\mathrm{E}|x-\xi_J|)^\alpha,\end{aligned} \tag{E.5}$$

by Jensen's or Hölder's inequality. Since $|\xi_J-J/k|\le k^{-1}$ and $\mathrm{E}(J/k)=x-x/k$,

$$\mathrm{E}|x-\xi_J|\le\frac{1}{k}+\mathrm{E}\Big|x-\frac{J}{k}\Big|\le\frac{1+x}{k}+\mathrm{E}\Big|\mathrm{E}\frac{J}{k}-\frac{J}{k}\Big|\le\frac{2}{k}+\sqrt{\mathrm{var}\,\frac{J}{k}}\le\frac{2}{k}+\frac{1}{2\sqrt{k}}.$$

The assertion of the lemma follows upon inserting this bound on the right side of (E.5).

If $\alpha\in(1,2]$, we can use Taylor's theorem with integral remainder to obtain

$$k\mathrm{E}\Big[F\Big(\frac{J+1}{k}\Big)-F\Big(\frac{J}{k}\Big)\Big]=\mathrm{E}f\Big(\frac{J}{k}\Big)+\frac{1}{k}\int_0^1 f'\Big(\frac{J}{k}+\frac{s}{k}\Big)(1-s)\,ds. \tag{E.6}$$

The second term on the right is bounded above by $(2k)^{-1}\|f'\|_\infty$. By the mean value theorem there exists for every t and h some $\xi\in[0,1]$ such that

$$\big|f(t+h)-f(t)-hf'(t)\big|=\big|hf'(t+\xi h)-hf'(t)\big|\le|h|L_{\alpha-1}(f')|\xi h|^{\alpha-1}\le L_{\alpha-1}(f')|h|^\alpha,$$

since f' is $(\alpha-1)$-smooth. Substituting $t=(k-1)x/k$ and $h=J/k-(k-1)x/k$, we find

$$\Big|f\Big(\frac{J}{k}\Big)-f\Big(\frac{k-1}{k}x\Big)-\Big(\frac{J}{k}-\frac{k-1}{k}x\Big)f'\Big(\frac{k-1}{k}x\Big)\Big|\le L_{\alpha-1}(f')\Big|\frac{J}{k}-\frac{k-1}{k}x\Big|^\alpha. \tag{E.7}$$

Hölder's inequality gives

$$\mathrm{E}\Big|\frac{J}{k}-\frac{k-1}{k}x\Big|^\alpha\le\Big(\mathrm{E}\Big|\frac{J}{k}-\frac{k-1}{k}x\Big|^2\Big)^{\alpha/2}=\Big(\mathrm{var}\,\frac{J}{k}\Big)^{\alpha/2}\le\Big(\frac{1}{4k}\Big)^{\alpha/2}. \tag{E.8}$$

Combining (E.4), (E.6), (E.7) and (E.8), and using that $\mathrm{E}(J/k)=(k-1)x/k$, we see that $\|f-b(\cdot;k,F)\|_\infty$ is bounded by

$$\begin{aligned}&\sup_{0\le x\le1}\Big|f(x)-f\Big(\frac{(k-1)x}{k}\Big)\Big|+\sup_{0\le x\le1}\Big|f\Big(\frac{(k-1)x}{k}\Big)-\mathrm{E}f\Big(\frac{J}{k}\Big)\Big|+\frac{1}{2k}\|f'\|_\infty\\&\quad\le L_1(f)\frac{1}{k}+L_{\alpha-1}(f')\Big(\frac{1}{4k}\Big)^{\alpha/2}+\frac{1}{2k}\|f'\|_\infty.\end{aligned}$$

For $\alpha\le2$ the middle term dominates in order. ☐

The approximation error of Bernstein polynomials does not compare well with other approximation techniques in terms of dimension. Splines, wavelets or general polynomials based on k terms achieve an approximation error of the order $k^{-\alpha}$ for α-smooth densities. This may be thought of as a price paid for the shape-preserving property of Bernstein polynomials. Their use of many similar and hence redundant terms translates into higher complexity, and may reduce the speed of posterior convergence, depending on the prior. This may be avoided by clumping together terms, thus reducing the dimension of the approximating linear combination of the basis without sacrificing the quality of approximation.

For $k = l^2$, $l \in \mathbb{N}$ and $w \in \mathbb{S}_l$, define the *coarsened Bernstein polynomial* as

$$\tilde{b}(x; l^2, w) = \sum_{i=1}^{l} w_i \, l^{-1} \sum_{j=(i-1)l+1}^{il} \mathrm{be}(x; j, k+1-j). \tag{E.9}$$

If $w_j = F((j-1)/l, j/l])$, we shall also write $\tilde{b}(x; l^2, w)$ as $\tilde{b}(x; l^2, F)$. Like the classical Bernstein polynomials, these functions are mixtures of beta-densities $\mathrm{be}(x; j, k+1-j)$, for $j = 1, \ldots, k$. However they are linear combinations of only l fixed functions, equal to the averages of the beta-densities in blocks of l consecutive elements. Although these functions have lower complexity, their accuracy of approximation is comparable with Bernstein polynomials of order k, as the following result shows.

Lemma E.4 *For $\alpha \in (0, 1]$*[1] *and $f \in \mathfrak{C}^\alpha[0, 1]$ we have $\|f - \tilde{b}(\cdot; k, F)\|_\infty \leq 6L_\alpha(f)k^{-\alpha/2}$, where $F(x) = \int_0^x f(u)\, du$.*

Proof Define a function ϕ_k by $\phi_k(y) = il$ if $(i-1)l \leq y < il$, for $i = 1, \ldots, l = \sqrt{k}$. It is easy to verify that, for $J \sim \mathrm{Bin}(k-1, x)$,

$$\tilde{b}(x; k, F) = l\, \mathrm{E}\Big[F\Big(\frac{\phi_k(J)}{k}\Big) - F\Big(\frac{\phi_k(J) - l}{k}\Big)\Big]. \tag{E.10}$$

In view of Lemma E.3 it suffices to show that $\|b(\cdot; k, F) - \tilde{b}(\cdot; k, F)\|_\infty \leq 3L_\alpha(f)k^{-\alpha/2}$. By the definition of ϕ_k it follows that $|y - (\phi_k(y) - l)| \leq l$ for every y, whence

$$\mathrm{E}\Big|\frac{J}{k} - \frac{\phi_k(J) - l}{k}\Big|^\alpha \leq k^{-\alpha/2}. \tag{E.11}$$

By Taylor's theorem with integral remainder,

$$k\Big[F\Big(\frac{J+1}{k}\Big) - F\Big(\frac{J}{k}\Big)\Big]f = \int_0^1 f\Big(\frac{J}{k} + \frac{s}{k}\Big)\, ds,$$
$$l\Big[F\Big(\frac{\phi_k(J)}{k}\Big) - F\Big(\frac{\phi_k(J) - l}{k}\Big)\Big] = \int_0^1 f\Big(\frac{\phi_k(J) - l}{k} + \frac{s}{l}\Big)\, ds.$$

By (E.4) and (E.10), for all $x \in [0, 1]$,

$$\begin{aligned}
\Big|b(x; k, F) - \tilde{b}(x; k, F)\Big| &\leq \int_0^1 \mathrm{E}\Big|f\Big(\frac{\phi_k(J) - l}{k} + \frac{s}{l}\Big) - f\Big(\frac{J}{k} + \frac{s}{k}\Big)\Big|\, ds \\
&\leq L_\alpha(f) \int_0^1 \mathrm{E}\Big|\frac{\phi_k(J) - l}{k} + \frac{s}{l} - \frac{J}{k} - \frac{s}{k}\Big|^\alpha\, ds \\
&\leq L_\alpha(f) \int_0^1 \Big[\Big|\frac{s}{k} + \frac{s}{l}\Big|^\alpha + \mathrm{E}\Big|\frac{J}{k} - \frac{\phi_k(J) - l}{k}\Big|^\alpha\Big]\, ds \leq 3k^{-\alpha/2},
\end{aligned}$$

where the last inequality holds because of (E.11). □

[1] For $\alpha \in (1, 2]$, the result is false. Indeed, coarsening may be counterproductive in this case; see Problem E.3.

E.2 Splines

A *spline* function on an interval $[a, b)$ in $\mathbb{R}$ is a piecewise polynomial function with a given level of global smoothness. More precisely, given $K+1$ *knots* $a = t_0 < t_1 < \cdots < t_K = b$, a function $f\colon [a, b] \to \mathbb{R}$ is a spline of *order* q if the restriction $f|_{[t_{k-1},t_k]}$ of f to any subinterval is a polynomial of degree at most $q-1$ and $f \in \mathfrak{C}^{q-2}(a, b]$ (provided that $q \geq 2$).

Splines form a linear subspace of the space of all *piecewise polynomials*, which are the functions f such that every restriction $f|_{[t_{k-1},t_k]}$ is a polynomial. A piecewise polynomial of order q is described by q free coefficients in every interval, and hence the linear space of all piecewise polynomials of order K has dimension Kq. Splines are restricted to have equal left and right derivatives of orders $0, 1, \ldots, q-2$ at every interior knot point, leading to $q-1$ linear constraints at each of the $K-1$ interior knot points $t_1, \ldots, t_{K-1}$, and hence to $(K-1)(q-1)$ constraints in total. Thus the dimension of the spline space of order q is $J = Kq - (K-1)(q-1) = q + K - 1$.

Splines occur naturally as solutions in interpolation problems, by their ability to approximate smooth functions. In statistics they appear naturally as the solution of the penalized least square regression problem:

$$\operatorname*{argmin}_f \left\{ \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int f^{(m)}(x)^2\, dx \right\}.$$

The solution can be shown to be a spline of order $2m-1$ with knots at the observations (and the begin and end of the domain); it is known as a *smoothing spline*. The common choice $m = 2$ leads to *cubic splines*.

A convenient basis for the space of splines is the set of *B-splines* $B_{0,q}, \ldots, B_{J-1,q}$. For the interval $[0, 1]$ these can be defined recursively, as follows. For $q = 1$ the functions $B_{0,1}, \ldots, B_{K-1,1}$ are simply indicator functions $\mathbb{1}\{t_j \leq x < t_{j+1}\}$, for $j = 0, 1, \ldots, K-1$, having discontinuities at interior knot points $t_1, \ldots, t_{K-1}$. For $q = 2$, the basis functions are *tent functions*

$$B_{j,2}(x) = \begin{cases} \frac{x-t_j}{t_{j+1}-t_j}, & x \in [t_j, t_{j+1}), \\ \frac{t_{j+2}-x}{t_{j+2}-t_{j+1}}, & x \in [t_{j+1}, t_{j+2}), \\ 0, & \text{otherwise}, \end{cases} \qquad j = -1, 0, \ldots, K-1.$$

Here by convention $t_{-1} = t_0 = 0$ and $t_K = t_{K+1} = 1$. These functions are continuous, but not differentiable at the interior knot points. For a general q, define an extended knot sequence, with 0 and 1 are deliberately repeated q times, as

$$\overbrace{0, 0, \ldots, 0}^{q\text{ times}}, t_1, \ldots, t_{K-1}, \overbrace{1, 1, \ldots, 1}^{q\text{ times}}.$$

Then B-splines of order q can be written in terms of lower order ones as

$$B_{j,q}(x) = \frac{x - t_j}{t_{j+q-1} - t_j} B_{j,q-1}(x) + \frac{t_{j+q} - x}{t_{j+q} - t_{j+1}} B_{j+1,q-1}(x), \qquad j = -q+1, \ldots, K-1.$$

Figure E.1 gives a visible impression of these functions.

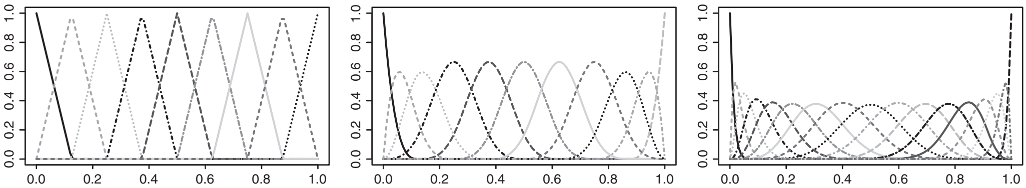

Figure E.1 B-spline basis functions of order 2, 4 and 10, for the knots at $0, 1/10, 2/10, \ldots, 1$ (i.e. $q = 2, 4, 10$ and $K = 10$).

From now on we assume that the order of the splines has been fixed at some value q. This will be dropped from the notation and the B-spline basis denoted by $B_1, B_2, \ldots, B_J$. Furthermore, for $\theta \in \mathbb{R}^J$ we write $\theta^\top B$ for the linear combination $\sum_{j=1}^J \theta_j B_j$.

For the half open unit interval $[0, 1)$, the uniform knot sequence k/K, for $k = 0, 1, \ldots, K$, is typically sufficient, but most of the following results are true more generally for "quasi-uniform" knot sequences, meaning that the ratio of the largest and smallest spacing between the knots is bounded by a universal constant. The following properties of B-splines are relevant for theoretical studies:

(i) $B_j \geq 0$, $j = 1, \ldots, J$,
(ii) $\sum_{j=1}^J B_j = 1$,
(iii) B_j is supported inside an interval of length q/K,
(iv) at most q functions B_j are nonzero at any given x.
(v) the integrals are given by

$$\int_0^1 B_j(x)\, dx = \begin{cases} j/(q(J-q+1)), & j = 1, 2, \ldots, q-1, \\ 1/(J-q+1), & j = q, q+1, \ldots, J-q+1, \\ (J-j+1)/(q(J-q+1)), & j = J-q+2, \ldots, J. \end{cases} \tag{E.12}$$

The first two properties express that the basis elements form a partition of unity, and the third and fourth properties mean that their supports are close to being disjoint if K is very large relative to q. The integrals of the B-spline functions located in the middle of the interval are identical as these functions are location shifts of each other. Only the most extreme B-spline functions have different integrals.

The following result shows that spline functions have excellent approximation properties for smooth functions.

Lemma E.5 *Suppose $q \geq \alpha > 0$. There exists a constant C depending only on q and α such that for every $f \in \mathfrak{C}^\alpha[0, 1]$ there exists $\theta \in \mathbb{R}^J$ with $\|\theta\|_\infty < \|f\|_{\mathfrak{C}^\alpha}$ such that*

$$\|\theta^\top B - f\|_\infty \leq C J^{-\alpha} \|f\|_{\mathfrak{C}^\alpha}.$$

Furthermore,

(a) *If f is strictly positive, for large J the vector θ in (a) can be chosen to have strictly positive coordinates.*[2]

[2] The condition that f is strictly positive is crucial. The best approximation $\theta^\top B$ with nonnegative coefficients θ of a nonnegative function f may have approximation error only $O(J^{-1})$ no matter how smooth f is; see De Boor and Daniel (1974).

(b) *If $0 < f < 1$, for large J the coordinates of the vector θ can be chosen between 0 and 1.*

(c) *If f is a probability density, then for large J there exists $\theta \in \mathbb{S}_J$ such that, for $B_j^* = B_j / \int_0^1 B_j(x)\,dx$,*

$$\|f - \theta^\top B^*\|_\infty \leq C J^{-\alpha} \|f\|_{\mathfrak{C}^\alpha}. \tag{E.13}$$

Proof The first part is well known in approximation theory and can be found in de Boor (1978), page 170.

For the proof of assertion (a), suppose that $f \geq \epsilon > 0$ pointwise. By Corollaries 4 and 6 in Chapter 11 of de Boor (1978), for each θ_i, there exists a constant C_1 that depends only on q such that $|\theta_j - c| \leq C_1 \sup_{x \in [t_{j+1}, t_{j+q-1}]} |f(x) - c|$, for any choice of the constant c. Choose c equal to the minimum of f over the interval $[t_{j+1}, t_{j+q-1}]$. If this minimum is attained at t^*, then $|f(x) - c| \leq C_2 |x - t^*|^{\min(\alpha,1)} \leq C_2 (q/J)^{\min(\alpha,1)}$, for every $x \in [t_{j+1}, t_{j+q-1}]$ and some constant $C_2 > 0$, since $f \in \mathfrak{C}^\alpha$. Then for $J > q(C_1 C_2/\epsilon)^{\max(1/\alpha,1)}$, we have $\theta_j > c - C_1 (q/J)^{\min(\alpha,1)} \geq 0$.

Part (b) follows from the preceding proof of (a) when also applied to $1 - f > 0$.

For part (c), we apply (b) to obtain existence of $\eta_1 \in (0,\infty)^J$ such that $\|f - \eta_1^\top B\| \lesssim J^{-\alpha}$. If $\eta_{2,i} = \eta_{1,i} \int_0^1 B_j(x)\,dx$, for $j = 1, \ldots, J$, then $\|f - \eta_2^\top B^*\|_\infty \lesssim J^{-\alpha}$, and in particular $\|\eta_2^\top B\|$ is bounded. By integration it follows that $|1 - \|\eta_2\|_1| = |1 - \sum_{j=1}^J \eta_{2,j}| \lesssim J^{-\alpha}$. Finally, if $\theta = \eta_2 / \|\eta_2\|_1$, then $\|f - \theta^\top B^*\|_\infty \leq \|f - \eta_2^\top B^*\|_\infty + \|\eta_2^\top B^*\|_\infty |1 - (\|\eta_2\|_1)^{-1}| \lesssim J^{-\alpha}$. □

Lemma E.6 *For any $\theta \in \mathbb{R}^J$, with $\|\cdot\|_2$ denoting the Euclidean and the $\mathbb{L}_2[0,1]$-norm,*

$$\|\theta\|_\infty \lesssim \|\theta^\top B\|_\infty \leq \|\theta\|_\infty, \qquad \|\theta\|_2 \lesssim \sqrt{J}\, \|\theta^\top B\|_2 \lesssim \|\theta\|_2.$$

Proof The first inequality is proved by de Boor (1978), page 156, Corollary XI.3). The second is immediate from the fact that the B-spline basis forms a partition of unity.

Let I_j be the interval $[(j-q)/K \vee 0, j/K \wedge 1]$. By Equation (2) on page 155 of de Boor (1978), we have

$$\sum_j \theta_j^2 \lesssim \sum_i \|\theta^\top B_{|I_j}\|_\infty^2 \lesssim \sum_j K \|\theta^\top B_{|I_j}\|_2^2.$$

The last inequality follows, because $\theta^\top B_{|I_j}$ consists of at most q polynomial pieces, each on an interval of length K^{-1}, and the supremum norm of a polynomial of order q on an interval of length L is bounded by $L^{-1/2}$ times the $\mathbb{L}_2$-norm, up to a constant depending on q. To see the third, observe that the squared $\mathbb{L}_2[0,1]$-norm of the polynomial $x \mapsto \sum_{j=0}^{q-1} \alpha_j x^j$ on $[0,1]$ be the quadratic form $\alpha^\top \mathrm{E}(U_q U_q^T)\alpha$ for $U_q = (1, U, \ldots, U^{q-1})$ and U a uniform $[0,1]$ variable. The second moment matrix $\mathrm{E}(U_q U_q^T)$ is nonsingular and hence the quadratic form is bounded below by a multiple of $\|\alpha\|^2$. This yields the third inequality.

By property (iii) of the B-spline basis, at most q elements $B_j(x)$ are nonzero for every given x, say for $j \in J(x)$. Therefore,

$$(\theta^\top B(x))^2 = \Big(\sum_{j\in J(x)} \theta_j B_j(x)\Big)^2 \leq \sum_{j\in J(x)} \theta_j^2 B_j^2(x)\, q, \tag{E.14}$$

by the Cauchy-Schwarz inequality. Since each B_j is supported on an interval of length proportional to $K^{-1} \asymp J^{-1}$ and takes its values in [0, 1], its $\mathbb{L}_2[0,1]$-norm is of the order $J^{-1/2}$. Combined with (E.14), this yields $\int_0^1 (\theta^\top B(x))^2\,dx \lesssim q\|\theta\|_2^2/J$, leading to the fourth inequality. $\square$

Smooth functions on the multi-dimensional domain $[0,1]^k$ may be approximated using a basis formed by tensor products of univariate splines:

$$B_{j_1\cdots j_k}(x_1,\ldots,x_k) = \prod_{l=1}^k B_{j_l}(x_l), \qquad 1 \leq j_l \leq J_l, \text{ for } l = 1,\ldots,k.$$

This collection of tensor-products inherits several useful properties of univariate B-splines, such as nonnegativity, adding to unity, being supported within a cube with sides of length q/K, and at most q^k many tensor-product B-splines being nonzero at any given point. The approximation properties also carry over.

For $\alpha = (\alpha_1,\ldots,\alpha_k) \in \mathbb{N}^k$, define the *anisotropic Hölder class* $\mathfrak{C}^\alpha([0,1]^k)$ to be the collection of all functions that are α_l times partially differentiable in their lth coordinate with partial derivative satisfying

$$\Big\| \frac{\partial^{\sum_{l=1}^k r_l} f}{\partial x_1^{r_1}\cdots \partial x_k^{r_k}} \Big\|_\infty < \infty, \qquad 0 \leq r_l \leq \alpha_l, \quad \text{for } l = 1,\ldots,k. \tag{E.15}$$

Lemma E.7 *There exists a constant $C > 0$ that depends only on q such that for every $f \in \mathfrak{C}^\alpha([0,1]^k)$ there exists $\theta = (\theta_{j_1\cdots j_k}\colon 1 \leq j_l \leq J_l), \in \mathbb{R}^{\prod_{l=1}^k J_l}$ with*

$$\|f - \theta^\top B\|_\infty \leq C \sum_{l=1}^k J_l^{-\alpha_l} \Big\| \frac{\partial^{\alpha_l} f}{\partial x_l^{\alpha_l}} \Big\|_\infty.$$

Furthermore,

(a) *If f is strictly positive, then every component of θ can be chosen to be positive.*
(b) *If $0 < f < 1$, then every component of θ can be taken to lie between 0 and 1.*
(c) *If f is a probability density function, the approximation order is maintained when replacing the tensor product splines by their normalized versions $B^*_{j_1\cdots,j_k}$ and restricting θ to $\mathbb{S}_{\prod_{l=1}^k J_l}$.*

If $\alpha_1 = \cdots = \alpha_k$, then these assertions extend to non-integer values of α and the isotropic Hölder class of order α (see Definition C.4).

Proof The first assertion is established in Theorem 12.7 in Schumaker (2007).

Assertion (a) can be proved using the arguments given in the proof of part (a) of Lemma E.5, once we show that for any $c > 0$,

$$|\theta_{j_1\cdots j_k} - c| \leq C_1 \max_l \sup_{x_l \in [t_{i+1,l}, t_{i+q-1,l}]} |f(x) - c|.$$

This requires bounding the absolute values of the coefficients using the values of the target function, which can be accomplished by using a dual-basis – a class of linear functionals which recovers values of the coefficients in a spline expansion from the values of the function; see Schumaker (2007). Clearly a dual basis for the multivariate B-splines is formed by tensor products of univariate dual bases and these can be chosen to be uniformly bounded; see Theorem 4.41 of Schumaker (2007). This bounds the maximum value of coefficients of spline approximations by a constant multiple of the $\mathfrak{L}_\infty$-norm of the target function. This gives the desired bound.

Proofs of parts (b) and (c) can follow the same arguments used to derive the respective parts of Lemma E.5. Finally, for the isotropic case, the approximation property is not restricted to the case of integer smoothness; see Schumaker (2007). The rest can be completed by repeating the same arguments. □

E.3 Wavelets

A drawback of polynomials and trigonometric functions is that they are globally defined, so that precise behavior of an approximation at a given point automatically has consequences for the whole domain. Spline functions overcome this drawback by partitioning the interval, and permit local modeling through the B-spline basis, but they lack the ease of an orthonormal basis. Wavelets provide both local modeling and orthogonality, and have excellent approximation properties relative to a wide range of norms. They come as a double-indexed basis ψ_{jk}, where j corresponds to a "resolution level," with bigger j focusing on smaller details, and k corresponds to location.

For a given function $g\colon \mathbb{R} \to \mathbb{R}$ and $j, k \in \mathbb{Z}$, let g_{jk} denote the function

$$g_{jk}(x) = 2^{j/2} g(2^j x - k).$$

Thus j gives a binary scaling (or "dilation") and k shifts the function. The leading scaling factor $2^{j/2}$ is chosen so that every g_{jk} has the same $\mathbb{L}_2(\mathbb{R})$-norm as g. For a given resolution j one needs 2^j functions g_{jk} to shift the function over a unit length interval.

The surprising fact is that there exist compactly supported functions $\psi\colon \mathbb{R} \to \mathbb{R}$, called *wavelet* or *mother wavelet*, such that $\{\psi_{jk}\colon j, k \in \mathbb{Z}\}$ are an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$, and, moreover, the projection of a function on the subset of functions with resolution up to a given level is an excellent approximation to the function. Furthermore, the contributions of the resolution levels j up to a given level can be expressed as an infinite series of orthonormal translates of another compactly supported function $\phi\colon \mathbb{R} \to \mathbb{R}$, called *scaling function* or *father wavelet*. Thus a wavelet expansion of a given function $f \in \mathbb{L}_2(\mathbb{R})$ (starting at level $j = 0$) takes the form

$$f(x) = \sum_{k \in \mathbb{Z}} \alpha_k \phi_{0k}(x) + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(x). \tag{E.16}$$

The functions ϕ_{0k}, ψ_{jk}, for $k \in \mathbb{Z}$ and $j = 0, 1, \ldots$ form an orthonormal basis of $\mathbb{L}_2(\mathbb{R})$.

These functions are said to give a *multi-resolution* decomposition. If V_0 is the space spanned by the functions ϕ_{0k}, for $k \in \mathbb{Z}$, and W_j is the space spanned by ψ_{jk}, for $k \in \mathbb{Z}$, then $V_0, W_0, W_1, \ldots$ are orthogonal and the spaces $V_j = V_0 \oplus W_0 \oplus \cdots \oplus W_j$ are a nested sequence $V_0 \subset V_1 \subset \cdots$ whose span is $\mathbb{L}_2(\mathbb{R})$. Rather than at $j = 0$ as in the preceding display, the expansion can be started at an arbitrary $j > 0$, corresponding to the decomposition $\mathbb{L}_2(\mathbb{R}) = V_j \oplus W_j \oplus W_{j+1} \oplus \cdots$. The dilations ϕ_{jk}, for $k \in \mathbb{Z}$, of the father wavelet are an orthonormal basis of V_j, for every j. Therefore the expansion (E.16) remains valid if the ϕ_{0k} are replaced by the ϕ_{Jk} (with adapted coefficients α_k) and simultaneously the sum over j is started at $j = J$ instead of at $j = 0$.

By orthonormality the coefficients in (E.16) can be calculated as the inner products $\alpha_k = \int f(x)\phi_{0k}(x)\,dx$ and $\beta_{jk} = \int f(x)\psi_{jk}(x)\,dx$, for $k \in \mathbb{Z}$ and $j = 0, 1, \ldots$ Furthermore, the $\mathbb{L}_2$-norm of the function f is simply the ℓ_2-norm of the set of all coefficients in the expansion. For sufficiently regular wavelets, $\mathbb{L}_r$-norms for $r \neq 2$ can also be related to ℓ_r-norms of the coefficients, but less perfectly. Within a resolution level there is equivalence up to constants: for $1 \leq r \leq \infty$ and $j = 0, 1, \ldots,$

$$\Big\|\sum_k \alpha_k \phi_{0k}\Big\|_r \asymp \Big(\sum_k |\alpha_k|^r\Big)^{1/r},$$
$$\Big\|\sum_k \beta_{jk}\psi_{jk}\Big\|_r \asymp 2^{j(1/2-1/r)}\Big(\sum_k |\beta_{jk}|^r\Big)^{1/r}. \tag{E.17}$$

The factors $2^{j(1/2-1/r)}$ may seem surprising, but they simply arise because the wavelets ψ_{jk} have been scaled so that they have the same $\mathbb{L}_2$-norm; the factor $2^{j(1/2-1/r)}$ corrects this for the $\mathbb{L}_r$-norm. In general the equivalence in the preceding display does not extend to the full representation (E.16), but together with the triangle inequality they can be used to obtain at least an upper bound on the $\mathbb{L}_r$-norm of a function f in terms of its wavelet coefficients.

An approximation to f is obtained by truncating the series in j at a given resolution level J. In fact, this gives the orthogonal projection onto the span of the functions ϕ_{0k}, ψ_{jk}, for $k \in \mathbb{Z}$ and $j \leq J$. The accuracy of this approximation is usually described in terms of Besov spaces, which are defined as follows.

For a function f and number h the first difference is defined by $\Delta_h f(x) = f(x+h) - f(x)$; higher-order differences are defined by repetition as $\Delta_h^{r+1} f = \Delta_h \Delta_h^r f$, for $r \in \mathbb{N}$, where $\Delta_h^0 f = f$. The modulus of order r of f is defined as

$$\omega_{r,p}(f;t) = \sup_{h:0<h<t} \|\Delta_h^r f\|_p.$$

If f is defined on an interval in $\mathbb{R}$, then the rth difference $\Delta_h^r f(x)$ is only defined when both x and $x + rh$ are contained in the interval. The $\mathbb{L}_p$-norm $\|\Delta_h^r f\|_p$ is then understood to be relative to the Lebesgue measure on the corresponding reduced domain.

Definition E.8 (Besov space) For $p, q \geq 1$ and $\alpha > 0$ the *Besov* (p, q, α) *seminorm* of a measurable function $f\colon \mathfrak{X} \to \mathbb{R}$ on a bounded or unbounded interval $\mathfrak{X} \subset \mathbb{R}$ is defined as

$$\|f\|_{p,q,\alpha} = \begin{cases} \|f\|_p + \big[\int_0^\infty (h^{-\alpha}\omega_{\bar{\alpha},p}(f;h))^q\, h^{-1}\,dh\big]^{1/q}, & 1 \leq q < \infty, \\ \|f\|_p + \sup_{0<h<1} h^{-s}\omega_{\bar{\alpha},p}(f;h), & q = \infty. \end{cases} \tag{E.18}$$

Here $\bar{\alpha}$ may be any integer that is strictly bigger than α. The corresponding *Besov space* $\mathfrak{B}^{\alpha}_{p,q}(\mathfrak{X})$ is defined as the set of all functions $f: \mathfrak{X} \to \mathbb{R}$ with $\|f\|_{p,q,\alpha} < \infty$.

This definition is intimidating, by referring to three parameters, difference operators of order bigger than α and a weighted integral. The parameter α is the level of smoothness of the functions; the pair (p, q) determines the way in which smoothness is measured. The norm becomes more stringent (the norm up to equivalence bigger; the Besov space smaller) if p or α are increased, or if q is decreased.

In the important special case that $p = q = \infty$, the Besov space with parameter α contains the Hölder space of order α (see Definition C.4), and it reduces to this space if $\alpha \notin \mathbb{N}$. Thus the Besov scale may be viewed as a refinement of the Hölder scale.

Another important special case is obtained for $p = q = 2$, when the Besov space reduces to the *Sobolev space* $\mathfrak{W}^{\alpha}(\mathfrak{X})$. For $\alpha \in \mathbb{N}$ this is set of functions whose αth derivative (in the distributional sense) is contained in $\mathbb{L}_2(\mathfrak{X})$; the norm $\|f\|_2 + \|f^{(\alpha)}\|_2$ is equivalent to the Besov norm. For noninteger values of α, the Sobolev space $\mathfrak{W}^{\alpha}(\mathbb{R})$ is defined as the set of all functions $f \in \mathbb{L}_2(\mathbb{R})$ with Fourier transform $\hat{f}$ satisfying $\int (1 + |\lambda|^{2\alpha})|\hat{f}(\lambda)|^2\, d\lambda < \infty$; the Sobolev norm is the square root of the latter integral and is equivalent to the Besov norm.

Besov spaces can also be described through wavelet expansions of the form (E.16). If the wavelets are sufficiently regular ("higher than α"; we skip the details), then the Besov norm of a function $f \in \mathfrak{B}_{p,q,\alpha}(\mathbb{R})$ is equivalent to a norm on the coefficients in its wavelet expansion, as follows:

$$\|f\|_{p,q,\alpha} \asymp \Big(\sum_{k\in\mathbb{Z}} |\alpha_k|^p\Big)^{1/p} + \Big[\sum_{j=0}^{\infty} 2^{j(\alpha+1/2-1/p)q}\Big(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\Big)^{q/p}\Big]^{1/q}, \qquad q < \infty,$$

$$\|f\|_{p,\infty,\alpha} \asymp \Big(\sum_{k\in\mathbb{Z}} |\alpha_k|^p\Big)^{1/p} + \sup_{j\geq 0}\Big[2^{j(\alpha+1/2-1/p)}\Big(\sum_{k\in\mathbb{Z}} |\beta_{jk}|^p\Big)^{1/p}\Big].$$

Thus a bound on the Besov norms implies a bound on the coefficients of its wavelet expansion, which in turn gives control on the approximation error.

Proposition E.9 *Let $f_{\leq J}$ be the orthogonal projection of the function f on the linear span up to resolution level J of a suitable wavelet basis of regularity higher than α. For any $1 \leq p, q \leq \infty$ and for $f \in \mathfrak{B}^{\alpha}_{p,q}(\mathbb{R})$ there exists a sequence $c(f) \in \ell_q$ with $\|c(f)\|_q \lesssim \|f\|_{p,q,\alpha}$ such that $\|f - f_{\leq J}\|_p \leq 2^{-J\alpha} c_J(f)$, for every J. In particular $\|f - f_{\leq J}\|_p \lesssim 2^{-J\alpha}\|f\|_{p,\infty,\alpha}$, for every J.*

Proof If $\|\beta_{j\cdot}\|_p$ is the ℓ_p-norm of the sequence $(\beta_{jk})_{k=1}^{\infty}$ of coefficients in the wavelet expansion of f, then $\|f - f_{\leq J}\|_p \lesssim \sum_{j>J} \|\beta_{j\cdot}\|_p 2^{j(1/2-1/p)}$, by the triangle inequality and (E.17). Furthermore, for f contained in the Besov space the ℓ_q-norm of the sequence $v_j = \|\beta_{j\cdot}\|_p 2^{j(\alpha+1/2-1/p)}$ is bounded above by $\|f\|_{p,q,\alpha}$. This readily gives the result in the case $q = \infty$. For finite q, we have $\sum_{j>J} \|\beta_{j\cdot}\|_p 2^{j(1/2-1/p)} = 2^{-J\alpha} c_J$, for $c_J = \sum_{j>J} v_j 2^{-(j-J)\alpha}$. By Hölder's inequality, with q' the conjugate of q,

$$|c_J|^q \leq \sum_{j>J} |v_j|^q 2^{-(j-J)\alpha q/2} \Big(\sum_{j>J} 2^{-(j-J)\alpha q'/2}\Big)^{q/q'}.$$

The second sum on the right is bounded by a constant, and hence $\sum_J |c_J|^q \leq \sum_J \sum_{j>J} |v_j|^q 2^{-(j-J)\alpha q/2}$, which can be seen to be bounded by a multiple of $\sum_j |v_j|^q$ with the help of Fubini's theorem. □

The Haar basis is a simple example of a wavelet basis. Its father and mother wavelet are $\phi = \mathbb{1}_{[0,1]}$ and $\psi = -\frac{1}{2}\mathbb{1}_{[0,\frac{1}{2})} + \frac{1}{2}\mathbb{1}_{[\frac{1}{2},1)}$, and the functions up to resolution level J are the functions that are constant on the dyadic intervals $(k2^{-J}, (k+1)2^{-J})$. Unfortunately, the discontinuity of these functions makes the Haar basis unsuitable for approximating functions of smoothness α higher than 1.

There are several popular wavelet bases of higher regularity, the most popular being *Daubechies wavelets*. No closed-form expressions of these functions exist, but many mathematical or statistical software packages contain a wavelet toolbox that allows numerical calculations.

Because a wavelet basis is defined through translation, its application to functions that are defined on a subinterval of $\mathbb{R}$ is awkward. One way around is to extend a given function to the full domain $\mathbb{R}$, and apply the preceding results to the extension. It can be shown that any function in a Besov space $\mathfrak{B}^\alpha_{p,q}[a,b]$ possesses such an extension of equal Besov norm, whence the Besov norm on a reduced interval can still be used to control approximation. If the interval $[a,b]$ is compact, then for each resolution level the support of only finitely many (of the order 2^j) of the translates ψ_{jk} will intersect the interval and hence the expansion can be reduced to finite sums in k.

Another solution is to adapt a given wavelet basis near the boundaries of the interval, so that orthonormality is retained within the interval. One construction (Cohen et al. 1993) gives for any smoothness level $\alpha > 0$ a wavelet basis $\{\psi_{jk}: 0 \leq k \leq 2^j - 1, j = 0, 1, \ldots\}$ for the interval [0, 1] with the following properties:

- $\mathrm{diam}(\mathrm{supp}(\psi_{jk})) \lesssim 2^{-j}$;
- $\|\psi_{jk}\|_\infty \lesssim 2^{j/2}$;
- $\#\{k': \mathrm{supp}(\psi_{j'k'}) \cap \mathrm{supp}(\psi_{jk}) \neq \varnothing\}$ is bounded by a universal constant, for any j and $j' \leq j$;
- $\#\{k': \mathrm{supp}(\psi_{j'k'}) \cap \mathrm{supp}(\psi_{jk}) \neq \varnothing\} \lesssim 2^{j'-j}$, for any j and $j' > j$;
- $\sum_{k=0}^{2^j-1} \|\psi_{jk}\|_\infty \lesssim 2^{j/2}$;
- $\int \psi_{jk}(x)\,dx = 0$, for all sufficiently large j;
- $g \in \mathfrak{B}^\alpha_{p,q}[0,1]$ if and only if $|\int g\psi_{jk}(x)\,dx| \lesssim 2^{-j(\alpha+1/2)}$.

E.4 Historical Notes

The results on polynomials and splines are classical; the results on wavelets date back to the 1980s and 90s. The books de Boor (1978), Cohen et al. (1993), DeVore and Lorentz (1993), Daubechies (1992), Schumaker (2007) and Zygmund (2002) contain most results and many references. For introductions from a statistical perspective, we refer to Härdle et al. (1998) and Giné and Nickl (2015).

Problems

E.1 Give a simpler proof of Lemma E.3 for the case $\alpha = 2$.

E.2 Let k be an odd integer. Show that the sum of the standard deviations of $\mathrm{Be}(\cdot; j, k+1-j)$, $j = 1, \ldots, k$, is $O(\sqrt{k})$. Hence for $k \to \infty$, there is increasing overlap in the terms of the mixture, especially for terms corresponding to j close to $k/2$. This motivates the idea behind Lemma E.4.

E.3 Let $f(x) = F'(x) = 2x$. Show that if k is a perfect square,

$$\begin{aligned}\tilde{b}(1; k, F) &= \frac{1}{\sqrt{k}}\Big[F(1) - F\Big(\frac{\sqrt{k}-1}{\sqrt{k}}\Big)\Big]\beta(1; k, 1) = \sqrt{k}\Big[F(1) - F\Big(\frac{k-1}{k}\Big)\Big] \\ &= 2 - \frac{1}{\sqrt{k}}.\end{aligned}$$

Thus the approximation error of the coarsened Bernstein polynomial is equal to $|f(1) - \tilde{b}(1; k, F)| = k^{-1/2}$, bigger than the approximation k^{-1} of the ordinary Bernstein polynomial.