Nonparametric Regression: Splines and RKHS Methods

Advanced Topics in Statistical Learning, Spring 2024

Ryan Tibshirani

Note: we're following the context, problem setup, notation, etc. from the last lecture.

1 Regression splines

Regression splines and smoothing splines are motivated from a different perspective than kernels and local polynomials; in the latter case, we started off with a special kind of local averaging, and moved our way up to a higher-order local models. With regression splines and smoothing splines, we build our estimator globally, from a set of select basis functions.

(We note that, at a broader level, the latter is often called the *synthesis* framework for modeling, where we build our estimator from a set of atoms—here being basis functions.)

1.1 Splines

These basis functions, as you might guess, are *splines*. Let's assume that $d = 1$. We'll stick to the univariate case for a little while, because splines are complex and interesting enough in dimension $d = 1$. A spline f of degree k with knots at $t_1 < \cdots < t_r$ is a piecewise polynomial of degree k that is continuous and has continuous derivatives of orders $1, \ldots, k-1$ at its knots. To be clear:

- f is a polynomial of degree k on each of $(-\infty, t_1], [t_1, t_2], \ldots, [t_r, \infty)$; and
- $D^\ell f$ is continuous at each of $t_1, \ldots, t_r$, for all $\ell = 0, \ldots, k-1$.

Splines have some special (some might say amazing) properties, and we will only really scratch the surface here. They have been a topic of interest among mathematicians and statisticians for a long time. Informally, a spline is a lot smoother than a piecewise polynomial, and so modeling with splines can serve as a way of reducing the variance of an estimator. See Figure 1 for an illustration.

A bit of statistical folklore: it is said that a cubic spline is so smooth, that one cannot detect the locations of its knots by eye!

1.2 Spline bases

How can we parametrize the space of k^{th} degree splines with knots at $t_1, \ldots, t_r$? The most natural way is to use the *truncated power basis*, $g_1, \ldots, g_{r+k+1}$, defined as

$$\begin{aligned} g_j(x) &= \frac{1}{(j-1)!} x^{j-1}, \quad j = 1, \ldots, k+1, \\ g_{j+k+1}(x) &= \frac{1}{k!} (x - t_j)_+^k, \quad j = 1, \ldots, r, \end{aligned} \tag{1}$$

Here $x_+ = \max\{x, 0\}$ denotes the positive part of x. From this we can see that the linear space of k^{th} degree splines with knots at $t_1, \ldots, t_r$ has dimension $r + k + 1$.

While (1) gives us a simple and natural basis, calculations involving truncated power bases functions can be troublesome because the ensuing basis matrix, $G \in \mathbb{R}^{n \times (r+k+1)}$ with entries

$$G_{ij} = g_j(x_i)$$

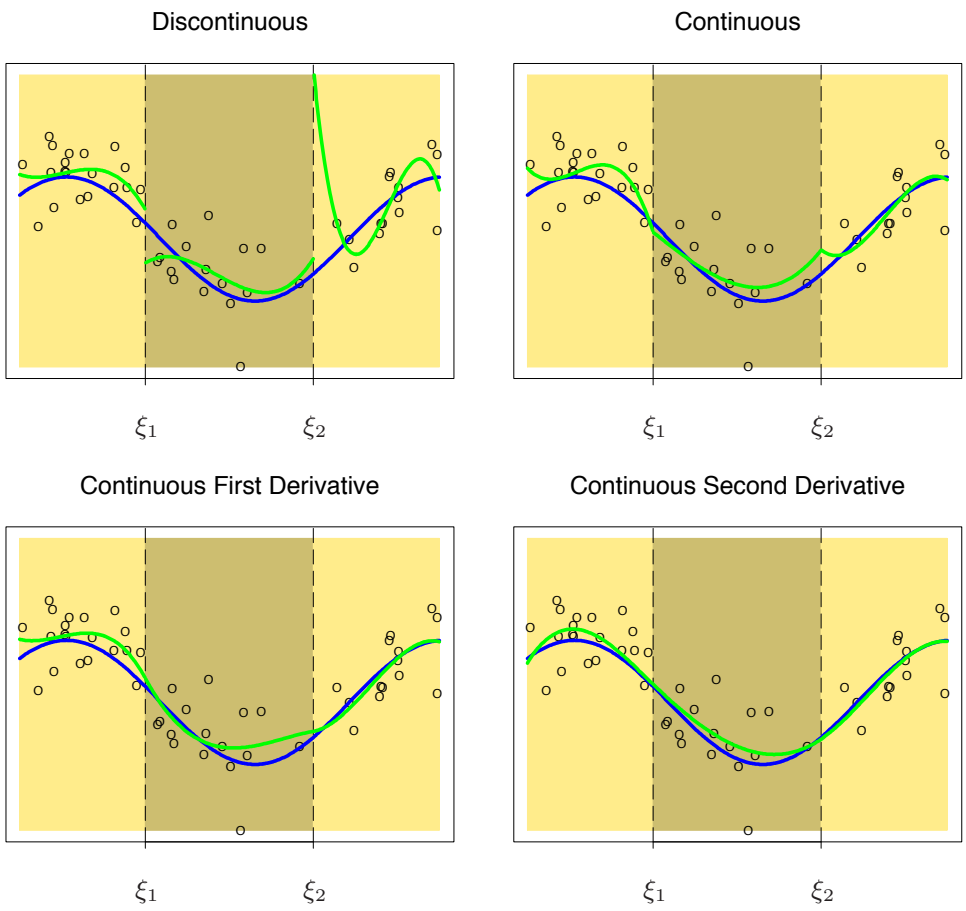

Figure 1: *Illustration of the effects of enforcing continuity at the knots for a cubic piecewise polynomial, and for various orders of the derivative. Credit: Chapter 5.2 of Hastie et al. (2009).*

is generally very poorly conditioned. A much better computational choice, both for speed and numerical stability, is called the *B-spline basis*. This was a major development in spline theory and is now much the standard in software. The key idea is that B-splines have local support: each B-spline basis function for a k^{th} degree spline space is supported on $k+2$ knots. Therefore the corresponding basis matrix is banded. It also tends to be much better conditioned.

Defining B-splines certainly requires more nuance than defining the truncated power basis, and the precise form of a B-spline is unimportant for the rest of this lecture. (If you are interested, you can find the details behind their construction in pretty much any standard reference on splines, such as de Boor (1978) or Schumaker (2007).)

1.3 Regress away!

A first idea: let's just perform regression on a spline basis. In other words, we use as our working model a k^{th} degree spline with knots at some pre-fixed locations $t_1, \ldots, t_r$. This means expressing f as

$$f = \sum_{j=1}^{r+k+1} \beta_j g_j$$

where $\beta_1, \ldots, \beta_{r+k+1} \in \mathbb{R}$ are coefficients and $g_1, \ldots, g_{r+k+1}$ is a basis for the space of k^{th} degree splines over the knots $t_1, \ldots, t_r$; for example, the truncated power basis or B-spline basis.

Letting $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ be the response vector, and $G \in \mathbb{R}^{n \times (r+k+1)}$ the basis matrix with entries

$$G_{ij} = g_j(x_i),$$

as before, we then just use least squares to determine estimate the coefficients, defining $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{r+k+1})$ to solve

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r+k+1} \beta_j g_j(x_i) \right)^2 \iff \underset{\beta}{\text{minimize}} \; \|Y - G\beta\|_2^2.$$

This yields the *regression spline* estimator, which makes predictions according to

$$\hat{f}(x) = \sum_{j=1}^{r+k+1} \hat{\beta}_j g_j(x).$$

Of course we know that $\hat{\beta} = (G^\mathsf{T} G)^{-1} G^\mathsf{T} Y$, so we can write this as

$$\hat{f}(x) = (g_1(x), \ldots, g_{r+k+1}(x))^\mathsf{T} (G^\mathsf{T} G)^{-1} G^\mathsf{T} Y = w(x)^\mathsf{T} Y,$$

which reveals that regression splines are linear smoothers.

This is a classic method, and can work well provided that we choose "good" knots $t_1, \ldots, t_r$; but in general choosing knots is a tricky business. There is a large literature on knot selection for regression splines via greedy methods like recursive partitioning. In practice, smoothing splines seem to be more popular, which we cover next.

2 Smoothing splines

Before delving into smoothing splines, we need to introduce a variant on the usual spline definition given above. To motivate it: a problem with spline estimates is that they can have somewhat erractive behavior—translating into high variance—at the boundaries of the input domain. (Recall that this is the opposite problem to that with kernel smoothing, which had poor bias at the boundaries.) This only gets worse as the polynomial order k gets larger.

2.1 Natural splines

One way to remedy this problem is to force the piecewise polynomial function to have a lower degree to the left of the leftmost knot, and to the right of the rightmost knot—this is exactly what *natural splines* do. A natural spline f of degree k with knots at $t_1 < \cdots < t_r$ is a piecewise polynomial of degree k such that:

- f is a polynomial of degree k on each of $[t_1, t_2], \ldots, [t_{r-1}, t_r]$;
- f is a polynomial of degree $(k-1)/2$ on $(-\infty, t_1]$ and $[t_r, \infty)$; and
- $D^\ell f$ is continuous at each of $t_1, \ldots, t_r$, for all $\ell = 0, \ldots, k-1$.

It is implicit here that natural splines are only defined for an odd degree k (linear, cubic, etc.) The choice $k = 3$ yields a natural cubic spline, by far the most common case: this is just a cubic spline that reduces to linear beyond the leftmost and rightmost knots.

What is the dimension of the span of k^{th} degree natural splines with knots at $t_1, \ldots, t_r$? Recall for splines, this was $r + k + 1$ (just count the number of truncated power basis functions). For natural splines, we can compute this dimension by counting as follows:

$$\underbrace{(k+1) \cdot (r-1)}_{a} + \underbrace{\left(\frac{(k-1)}{2} + 1 \right) \cdot 2}_{b} - \underbrace{k \cdot r}_{c} = r.$$

In the above:

- a is the number of parameters in the interior intervals $[t_1, t_2], \ldots, [t_{r-1}, t_r]$;
- b is the number of parameters in the exterior intervals $(-\infty, t_1], [t_r, \infty)$; and
- c is the number of constraints at the knots $t_1, \ldots, t_r$.

The fact that the total dimension is r is pretty remarkable; this is independent of k!

We note that there are simple modifications the truncated power basis that gives rise to a basis for natural splines, and similarly a modification of the B-spline basis for natural splines. And again, B-splines are the preferred parametrization for computational speed and stability.

2.2 Smooth away!

Smoothing splines, at the end of the day, are given by an ℓ_2-regularized regression over a natural spline basis that places knots at all input points $x_1, \ldots, x_n$. They circumvent the problem of knot selection because they just use all inputs as knots, and they make this possible—starting with a saturated model and producing meaningful function estimates—by using regularization to shrink the coefficients in the basis expansion.

Interestingly, we can motivate and define a smoothing spline directly from a functional minimization perspective. With input points $x_1, \ldots, x_n$ lying in an interval $[a, b]$, the *smoothing spline* estimator $\hat{f}$, of a given odd integer order $k \geq 1$, is defined to solve

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b [D^m f(x)]^2 \, dx, \tag{2}$$

for $m = (k+1)/2$. This is an infinite-dimensional optimization problem over all functions f for the which the criterion is well-defined and finite. The criterion in (2) trades off the squared error of f over (x_i, y_i), $i = 1, \ldots, n$, with a penalty term that is large when the order $(m-1)$ derivative of f is wiggly. The tuning parameter $\lambda \geq 0$ governs the tradeoff between these two terms.

By far the most commonly considered case is $k = 3$, the cubic smoothing spline, defined via

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b [f''(x)]^2 \, dx. \tag{3}$$

This is so common that often you'll just hear this called *the* smoothing spline.

2.3 Representer theorem

Remarkably, it so happens that the minimizer in the general smoothing spline problem (2) is unique, and it is a natural k^{th} degree spline with knots at the input points $x_1, \ldots, x_n$! This is known as a *representer theorem* for (2) (we will see more such results later).

Proof: we'll just consider the cubic case, $k = 3$, and follow Chapter 2.2 of Green and Silverman (1993) who give a nice direct proof.

The key result can be stated as follows: if g is any function on $[a, b]$ such that the penalty is well-defined (it has two derivatives, its second derivative is square integrable), and $x_1, \ldots, x_n \in [a, b]$ are arbitrary, then there exists a natural cubic spline f with knots at $x_1, \ldots, x_n$ such that:

- $f(x_i) = g(x_i)$, $i = 1, \ldots, n$; and
- $\int_a^b [f''(x)]^2 \, dx \leq \int_a^b [g''(x)]^2 \, dx$.

This would imply that we can restrict our attention in problem (3) to natural splines with knots at $x_1, \ldots, x_n$.

To prove the key result, we start with the fact that the cubic natural spline space with knots at $x_1, \ldots, x_n$ is n-dimensional, so given any n points $z_i = g(x_i)$, $i = 1, \ldots, n$, we can always find a natural spline f with

knots at $x_1, \ldots, x_n$ that satisfies $f(x_i) = z_i$, $i = 1, \ldots, n$. Now define $h = g - f$. Consider

$$\begin{aligned}
\int_a^b f''(x)h''(x)\,dx &= f''(x)h'(x)\Big|_a^b - \int_a^b f'''(x)h'(x)\,dx \\
&= -\int_{x_1}^{x_n} f'''(x)h'(x)\,dx \\
&= -\sum_{i=1}^{n-1} f'''(x)h(x)\Big|_{x_i}^{x_{i+1}} + \int_{x_1}^{x_n} D^4 f(x)h(x)\,dx \\
&= -\sum_{i=1}^{n-1} f'''(x_i^+)(h(x_{i+1}) - h(x_i)).
\end{aligned}$$

In the first line we used integration by parts; in the second line we used the fact that $f''(a) = f''(b) = 0$ and $f'''(x) = 0$ for $x \leq x_1$ and $x \geq x_n$, since f is a natural spline; in the third line we used integration by parts again; in the fourth we used the fact that f''' is constant on each open interval (x_i, x_{i+1}), and that $D^4 f = 0$, again because f is a natural spline. Since each $h(x_i) = 0$, we conclude from the last display that

$$\int_a^b f''(x)h''(x)\,dx = 0.$$

From this, it follows that

$$\int_a^b [g''(x)]^2\,dx = \int_a^b \big[f''(x) + h''(x)\big]^2\,dx = \int_a^b [f''(x)]^2\,dx + \int_a^b [h''(x)]^2\,dx$$

since the cross term is zero, and therefore

$$\int_a^b [f''(x)]^2\,dx \leq \int_a^b [g''(x)]^2\,dx,$$

with equality if and only if $h''(x) = 0$ for all $x \in [a, b]$. Note that $h'' = 0$ implies that h must be linear, and since we already know that $h(x_i) = 0$ for all $i = 1, \ldots, n$, this is equivalent to $h = 0$. In other words, the last display holds strictly except when $g = f$, so the solution in (3) is uniquely a natural spline with knots at the inputs.

2.4 Finite-dimensional form

From the representer result, we can choose a basis $\eta_1, \ldots, \eta_n$ for the set of k^{th} degree natural splines with knots at $x_1, \ldots, x_n$, and reparametrize the problem (2) as

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^n \left(y_i - \sum_{j=1}^n \beta_j \eta_j(x_i)\right)^2 + \lambda \int_a^b \left(\sum_{j=1}^n \beta_j D^m \eta_j(x)\right)^2 dx. \tag{4}$$

This is a finite-dimensional problem, and after we solve for the coefficients $\hat{\beta} \in \mathbb{R}^n$, the smoothing spline estimator is simply given by

$$\hat{f}(x) = \sum_{j=1}^n \hat{\beta}_j \eta_j(x).$$

Defining the basis matrix $N \in \mathbb{R}^{n \times n}$ and penalty matrix $\Omega \in \mathbb{R}^{n \times n}$ to have entries

$$N_{ij} = \eta_j(x_i), \quad \Omega_{ij} = \int_a^b D^m \eta_i(x) D^m \eta_j(x)\,dx,$$

the problem in (4) can be written more succintly as

$$\underset{\beta}{\text{minimize}} \ \|Y - N\beta\|_2^2 + \lambda \beta^\mathsf{T} \Omega \beta, \tag{5}$$

which a type of generalized ridge regression problem. From (5), it is not hard to see that the solution has the explicit form $\hat{\beta} = (N^T N + \lambda\Omega)^{-1} N^T Y$, and thus, we can write the smoothing spline as

$$\hat{f}(x) = (\eta_1(x), \ldots, \eta_n(x))^T (N^T N + \lambda\Omega)^{-1} N^T Y = w(x)^T Y,$$

which means, once again, smoothing splines are linear smoothers.

A remark on computation: the coefficients $\hat{\beta} = (N^T N + \lambda\Omega)^{-1} N^T Y$ can be computed in $O(n)$ operations. For this, we form N using the B-spline basis (for natural splines), since then the matrix $N^T N + \Omega I$ will be banded. In fact, more specialized computations are possible by taking advantage of more precise structure (beyond bandedness) afforded by splines. Altogether, in practice, smoothing spline computations are extremely fast.

2.5 Equivalent kernel

Recall that we can write a smoothing spline prediction as $\hat{f}(x) = w(x)^T Y$, for a weight function $w(x) = (w_1(x), \ldots, w_n(x)) \in \mathbb{R}^n$. We know the analytic form of this weight function, but how about its qualitative behavior? To be more precise, if we denote each component function by $w_i(x) = w(x, x_i)$ to emphasize that this weight gets attributed to (x_i, y_i) in the weighted sum $w(x)^T Y = \sum_{i=1}^n w(x, x_i) y_i$ which gives us the smoothing spline prediction at x, then we can ask the following questions:

- What shape does $z \mapsto w(x, z)$ have? Is it kernel-like?
- Does this shape change as we vary the test point x?

It's easy to just read off the answers to these questions from the rows of the smoother matrix $S = N(N^T N + \lambda\Omega)^{-1} N^T$, since in our expanded notation, its elements are

$$S_{ij} = w_j(x_i) = w(x_i, x_j).$$

Now, something very interesting happens when we plot a few rows of S. For evenly-spaced inputs, they look like the translations of the same kernel: see the left panel of Figure 2. Even more interestingly, for unevenly-spaced inputs, the rows still have a kernel shape but now the bandwidth appears to adapt to the density of the input points: lower density, larger bandwidth. See the right panel of Figure 2.

What we are seeing is an empirical validation of a beautiful asymptotic result by Silverman (1984), who proved that the cubic smoothing spline estimator is asymptotically equivalent to a kernel regression estimator, with an unusual choice of kernel. Specifically, under suitable regularity conditions, we have the large n approximation,[1]

$$w(x, z) \approx \frac{1}{h(x)} \frac{1}{p(x)} K\left(\frac{x - z}{h(x)}\right),$$

where K is the "Silverman kernel":

$$K(t) = \frac{1}{2} \exp(-|t|/\sqrt{2}) \sin(|t|/\sqrt{2} + \pi/4),$$

and we have the local bandwidth:

$$h(x) = \left[\frac{\lambda}{p(x)}\right]^{1/4},$$

where $p(x)$ is the density of the input distribution at x. That is, the bandwidth adapts to the local distribution of inputs.

The Silverman kernel is "kind of" a higher-order kernel. It satisfies

$$\int K(t)\, dt = 1, \quad \int t^j K(t)\, dt = 0,\ j = 1, \ldots, 3, \quad \text{but} \quad \int t^4 K(t)\, dt = -24.$$

So it lies outside the scope of usual kernel analysis. There is a lot more work building off of Silverman's initial work that connects smoothing splines to equivalent kernels.

[1] Silverman actually shows that $w(x, z) \approx \frac{1}{h(z)} \frac{1}{p(z)} K(\frac{x-z}{h(z)})$, but the version we are stating has a more natural interpretation with respect to Figure 2, and is a consequence of symmetry of the smoother matrix S.

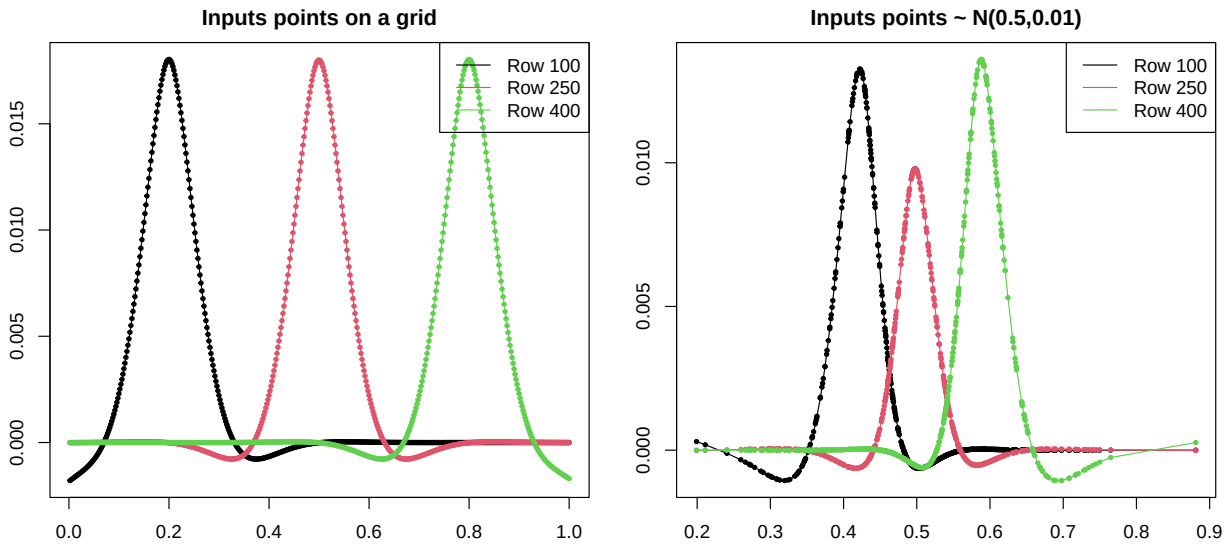

Figure 2: *Three sample rows of the cubic smoothing spline operator S defined over $n = 500$. Left: evenly-spaced inputs on a grid. The weights look like they are given by translations of the same kernel. Right: inputs drawn i.i.d. from $N(0.5, 0.01)$. The weights still look like kernels, but the bandwidth is now larger in low-density regions of the input domain.*

2.6 Rate of convergence

Compared to that for kNN and kernel smoothing, the error analysis for smoothing splines is more nuanced. Hence we'll dedicate a whole lecture to learning the tools behind it, a bit later in the course. The upshot is that we'll learn a general framework, and accompanying probabilistic tools, that can be used to analyze numerous other nonparametric estimators (such as other penalized empirical risk minimizers).

The punchline for smoothing splines will be as follows. If we assume that $f_0 \in W^{m,2}(L; [0,1])$ for a constant $L > 0$, we write to mean that $D^m f_0$ exists in the weak sense (to be defined precisely below) and $\int_0^1 [D^m f_0(x)]^2 \, dx \leq L^2$, and we assume mild conditions on the input points and sub-Gaussian noise, then the smoothing spline estimator of degree $k = 2m - 1$ with tuning parameter $\lambda \asymp n^{1/(2m+1)}$ satisfies

$$\max \left\{ \|\hat{f} - f_0\|_n^2, \, \|\hat{f} - f_0\|_2^2 \right\} \lesssim n^{-2m/(2m+1)} \quad \text{in probability},$$

with respect to randomness in the draws of (x_i, y_i), $i = 1, \ldots, n$.

Here, recall, we are abbreviating the population and empirical L^2 norms based on a test sample $x_0 \sim P$ and training samples $x_1 \ldots, x_n \sim P_n$, respectively, by

$$\|f\|_2^2 = \mathbb{E}[f^2(x_0)] = \int f^2(x) \, dP(x), \quad \text{and} \quad \|f\|_n^2 = \frac{1}{n} \sum_{i=1}^n f^2(x_i),$$

and we are using "$A_n \lesssim b_n$ in probability" to mean that $A_n = O_p(b_n)$.

3 Sobolev theory*

In somewhat of an interlude, we use this section to introduce Sobolev spaces. This material can be mostly skipped without interrupting the flow of understanding the main ideas in the rest of this lecture, hence the asterisk. Well actually, you should probably read the definition of weak differentiability, but you can treat the rest as optional. Yes, optional, but also super interesting and fairly fundamental in many ways, which is why we include some of the core details of Sobolev spaces here. To learn more, an excellent reference is Evans (2010).

3.1 Weak derivatives

First we introduce what is known as *weak differentiability*. A function $f : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^d$ is an open set, is called weakly differentiable provided that there is some function $g : U \to \mathbb{R}^d$ such that

$$\int_U f(x) D\phi(x)\, dx = -\int_U g(x)\phi(x)\, dx, \quad \text{for all } \phi \in C_c^\infty(U), \tag{6}$$

where $C_c^\infty(U)$ is the set of all infinitely differentiable functions with compact support in U, and $D\phi$ is the derivative of one such function ϕ. The function g satisfying (6) is unique almost everywhere, and denoted by Df henceforth, called the weak derivative of f.

The motivation for this definition of weak differentiability is integration by parts: if f, ϕ are differentiable in the classical (usual) sense, then $\int_U f(x)D\phi(x)\, dx = -\int_U Df(x)\phi(x)\, dx$ by integration by parts (applied componentwise), where there are no boundary terms because ϕ has compact support. Thus, we define the weak derivative $g = Df$ such that this holds for all infinitely differentiable compactly-supported ϕ.

Higher-order derivatives follow similarly. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$, recall, we write $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and we denote

$$D^\alpha \phi = \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_d^{\alpha_d}},$$

interpreted in the classical (usual) sense for a function ϕ. In the weak sense, f is said to have α^{th} weak differentiable g if

$$\int_U f(x) D^\alpha \phi(x)\, dx = -(1)^{|\alpha|} \int_U g(x)\phi(x)\, dx, \quad \text{for all } \phi \in C_c^\infty(U). \tag{7}$$

Again, the function g satisfying (7) is unique almost everywhere, and we denote by $D^\alpha f$.

Of course, the weak derivative reduces to the classical derivative exists when the latter exists. However, weak differentiability is more general. An example of a function that is weakly differentiable but not classically differentiable is

$$f(x) = x_+, \quad \text{with weak derivative } Df(x) = 1\{x > 0\}.$$

An example of a function that is *not* weakly differentiable is $f(x) = 1\{x > 0\}$. For this, one can check that the condition (6) cannot hold: taking ϕ to be any function whose support includes 0, we learn that

$$\phi(0) = \int g(x)\phi(x)\, dx.$$

This cannot possibly be true for all infinitely differentiable ϕ. This cannot possibly be true for all ϕ unless $g = 0$. But if $g = 0$, then we cannot recover $\phi(0)$ through integration, since the right-hand side above will always be 0.

3.2 Sobolev spaces

We are now equipped to define what are called *Sobolev spaces*. For an integer $k \geq 0$, and $1 \leq p \leq \infty$, and an open domain $U \subseteq \mathbb{R}^d$, we define the Sobolev space

$$W^{k,p}(U) = \Big\{ f : U \to \mathbb{R} \,:\, D^\alpha f \text{ exists in the weak sense, and } \|D^\alpha f\|_p < \infty \text{ for all } |\alpha| \leq k \Big\}.$$

Here $\|\cdot\|_p$ denotes the L^p norm on U:

$$\|f\|_p^p = \int_U |f(x)|^p\, dx \quad \text{for } p < \infty,$$

and $\|f\|_\infty = \operatorname{ess\,sup}_{x \in U} |f(x)|$ (the essential supremum, the smallest upper bound possible over all subsets of U of full measure).

Sobolev spaces are central in the study of partial differential equations, which is commonly where you'll find people (mathematicians) referring to them. But they also play an important role in nonparametric statistics. For example, if you recall the smoothing spline problem (2), where we said that "the minimization is all functions f for the which the criterion is well-defined and finite", we can take the domain here to be the L^2 Sobolev space[2]

$$W^{m,2}([a,b]) = \Big\{ f : [a,b] \to \mathbb{R} \,:\, D^\alpha f \text{ exists in the weak sense, and } \|D^\alpha f\|_2^2 < \infty, \text{ for all } \alpha \leq m \Big\}.$$

We note that, generally speaking, when working in Sobolev spaces (just like L^p spaces), we identify two functions that are equal almost everywhere. In other words, each element f in a Sobolev space is really an equivalence class of functions, any pair of which agree except on a set of measure zero.

Now we recall *Hölder spaces*. We touched on these in the lecture on kernels. Formally, for an integer $r \geq 0$, and $0 < \gamma \leq 1$, we define the Hölder space

$$C^{r+\gamma}(U) = \Bigg\{ f : U \to \mathbb{R} \,:\, D^\alpha f \text{ exists, } \|D^\alpha f\|_\infty < \infty \text{ for all } |\alpha| \leq r, \text{ and } \sup_{x \neq z} \frac{|D^\alpha f(x) - D^\alpha f(z)|}{\|x - z\|_2^\gamma} < \infty \text{ for all } |\alpha| = r \Bigg\}.$$

In the above, we can interpret $D^\alpha f$ as a classical derivative, and $\|\cdot\|_\infty$ as the supremum norm (we don't need the essential supremum norm). This may all seem more restrictive than Sobolev spaces, but as we'll discuss just below, we can actually identify the Hölder space with a particular Sobolev space.

Let's pause to examine the relationship between the Sobolev space $W^{k,\infty}(U)$ and the Hölder space $C^k(U)$. The former contains all functions f such that the α^{th} weak derivative of f is bounded for all $|\alpha| \leq k$. The latter contains all functions f such that the α^{th} derivative of f is bounded for all $|\alpha| < k$, and $D^\alpha f$ Lipschitz for $|\alpha| = k$. It turns out that these two conditions are really saying the same thing. Trying to make this statement precise for general domains is probably not worth it, so we'll just say that for a bounded domain U with smooth boundary, we can identify $W^{k,\infty}(U)$ with $C^k(U)$. To be precise, in doing so, we identify each f in the Sobolev space with its classically differentiable version.

Much more can be said about the connection between Sobolev and Hölder spaces, which is covered next.

3.3 Embedding theorems

Among the whole optional part of material on Sobolev spaces, this next bit really is the *most* optional, but it's too cool not to cover. We start by defining norms on Sobolev and Hölder spaces; namely,

$$\|f\|_{W^{k,p}(U)} = \begin{cases} \Big(\sum_{|\alpha| \leq k} \|D^\alpha f\|_p^p \Big)^{1/p} & \text{if } p < \infty \\ \sum_{|\alpha| \leq k} \|D^\alpha f\|_\infty & \text{if } p = \infty \end{cases} \qquad \text{for } f \in W^{k,p}(U),$$

and

$$\|f\|_{C^{r+\gamma}(U)} = \sum_{|\alpha| \leq r} \|D^\alpha f\|_\infty + \sum_{|\alpha| = r} \sup_{x \neq z} \frac{|D^\alpha f(x) - D^\alpha f(z)|}{\|x - z\|_2^\gamma} \qquad \text{for } f \in C^{r+\gamma}(U).$$

Equipped with these norms, $W^{k,p}(U)$ and $C^{r+\gamma}(U)$ are Banach spaces (complete normed linear spaces). In fact, when $p = 2$, the space $W^{k,p}(U)$ is a Hilbert space (complete inner product space) under the inner product $\langle f, g \rangle_{W^{k,p}(U)} = \sum_{|\alpha| \leq k} \langle D^\alpha f, D^\alpha g \rangle_2 = \sum_{|\alpha| \leq k} \int_U D^\alpha f(x) D^\alpha g(x)\, dx$.

[2]We'll ignore here, and in several other places, the fact that the domain is closed. Recall, we defined Sobolev spaces on open domains, and we will do the same for Hölder spaces as well. Hence, we are being slightly imprecise by abruptly allowing closed domains but it's easiest to do so for simplicity of exposition.

Now here comes the cool part, which are special inequalities[3] involving Sobolev norms that lead to what are known as *embedding theorems*. There are really several embedding theorems that were developed over many years, by different authors (Gagliardo, Nirenberg, Sobolev, Morrey, others), but now the totality of them is usually just called "the Sobolev embedding theorem". In a nutshell, this is what it says (recall d is the dimension of the domain U; and we're intentionally vague about what conditions it needs to satisfy).

The case $pk < d$, subcritical regime. In this case we a have lower smoothness-to-dimension ratio. Let $0 \leq \ell < k$ be an integer, and $p < q < \infty$, such that the pair satisfies

$$\frac{1}{q} - \frac{\ell}{d} = \frac{1}{p} - \frac{k}{d}.$$

Then we have, for a constant $C > 0$ depending only on k, p, ℓ, q and U,

$$\|f\|_{W^{\ell,q}(U)} \leq C\|f\|_{W^{k,p}(U)}, \quad \text{for all } f \in W^{k,p}(U).$$

This means that $W^{k,p}(U) \subseteq W^{\ell,q}(U)$, and moreover we have what is known as a *continuous embedding* of normed spaces. From the above inequality, we learn that any norm ball of radius ρ in the former space is contained in a normal ball of the latter, whose radius is at most $C\rho$.

Note that in the special case where we take $\ell = 0$, we get that for

$$q = \frac{pd}{d - pk},$$

and a constant $C > 0$ depending only on k, p, q and U, it holds that

$$\|f\|_q \leq C\|f\|_{W^{k,p}(U)}, \quad \text{for all } f \in W^{k,p}(U).$$

This means that $W^{k,p}(U) \subseteq L^q(U)$, and the embedding is continuous.

The case $pk > d$, supercritical regime. In this case we a have higher smoothness-to-dimension ratio. Let $0 \leq r < k$ be an integer, and $0 < \tau \leq 1$, such that the pair satisfies

$$\frac{r + \tau}{d} = \frac{k}{d} - \frac{1}{p}.$$

A technical detail: if the value of r, τ satisfying the above results in $\tau = 1$, then we need to "downgrade" it to a value less than 1. That is, let $\gamma = \tau$ if $\tau < 1$, and otherwise let $0 < \gamma < 1$. Then we have, for a constant $C > 0$ depending only on k, p, r, γ and U,

$$\|f\|_{C^{r+\gamma}(U)} \leq C\|f\|_{W^{k,p}(U)}, \quad \text{for all } f \in W^{k,p}(U).$$

This means that $W^{k,p}(U) \subseteq C^{r+\gamma}(U)$, and the embedding is continuous. In view of this, we can always identify each function $f \in W^{k,p}(U)$ with its r times classically differentiable version.

(The critical regime $pk = d$ is actually a bit more subtle and we'll skip over it.)

How do we interpret the Sobolev embedding theorem? In the subcritical regime, it says that we can trade off smoothness in the differential sense with smoothness in the integral sense. Given a k times (weakly) differentiable function, whose derivatives are in $L^p(U)$, the embedding theorem tells us that its derivatives of order $\ell < k$ are less "peaky" and more "evenly spread out", since they are $L^q(U)$ with $q > p$. In the supercritical regime, the conclusion is arguably even more fascinating—it says that knowing a function's derivatives are smooth in a global (integrated) sense tells us something about local smoothness of lower order derivatives, at each point in the domain. Given a k times (weakly) differentiable function, the embedding theorem starts off with a statement about the size of $\sum_{|\alpha| \leq k} \int_U |D^\alpha f(x)|^p \, dx$, and translates this into a statement about the Hölder constant of $D^\alpha f$ for $|\alpha| < k$. Figure 3 gives an illustration of the tradeoffs navigated by the Sobolev embedding theorem (these types of plots are sometimes called "DeVore diagrams" in honor of Ron DeVore).

[3] In an unfortunate clash of nomenclature with statistics, you'll often hear mathematicians calling these "estimates".

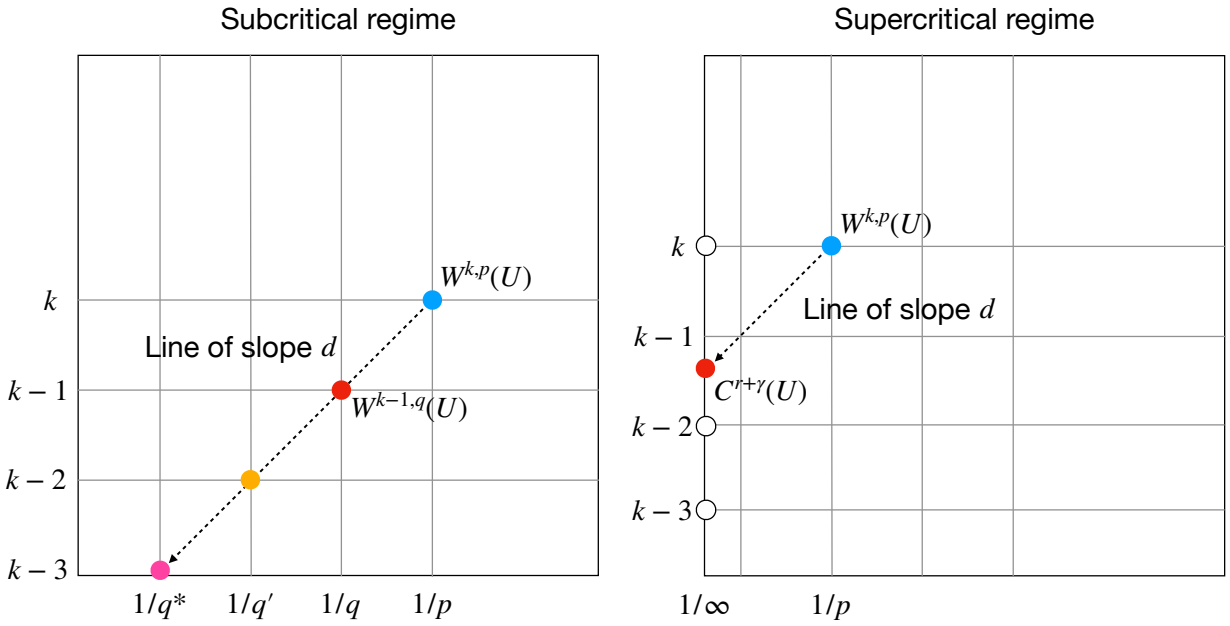

Figure 3: *Illustrations of the Sobolev embedding theorem. In each plot, we parametrize the x-axis by $1/p$ and the y-axis by k. Hence, as we move right-to-left, we increase the index of the L^p norm that is used to measure integral smoothness, and functions get less "locally peaked" and more "evenly spread out". As we move top-to-bottom, we decrease the number of (weak) derivatives under consideration. Left: subcritical regime with $pk < d$. We traverse a line of slope d, moving down and to the left, in order to visit the different Sobolev spaces into which $W^{k,p}(U)$ embeds, culminating in an intersection with the x-axis ($k = 0$) at $q = pd/(d - pk)$. Right: supercritical regime with $pk > d$. Similarly, we traverse a line of slope d to find the Hölder space into which $W^{k,p}(U)$ embeds, which is given by an intersection with the y-axis ($q = \infty$). The open white circles denote cases in which the embedding fails for $\gamma = 1$.*

4 Multivariate splines

We move on to methods for fitting multivariate splines in nonparametric regression. In a sense, both of the following statements are true.

1. There are several multivariate extensions of spline estimators, e.g, tensor product and thin plate splines (among others). We'll cover these.
2. There are no "real" multivariate extensions of spline estimators, and tensor product and thin plate splines are not "really" splines. In fact, even defining a multivariate spline is generally very tricky.

How can this be? Let's go about this backwards, and start by explaining the very last part of the second point: suppose you were to (reasonably) insist that a multivariate spline should be, like a univariate spline, a piecewise polynomial of a degree k that is C^{k-1}, which means it has continuous derivatives of all orders less than k. Then in general this is going to be tricky to fulfill when $d \geq 2$. This is essentially due to the number in the number of constraints such continuity imposes in the multivariate setting. Recall that a k^{th} degree polynomial in d dimensions is of the form (using multi-index notation $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$):

$$f(x) = \sum_{|\alpha| \leq k} \beta_\alpha \, x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}, \tag{8}$$

for coefficients β_α, $|\alpha| \leq k$. The number of coefficients, and hence the dimension of the space of k^{th} degree

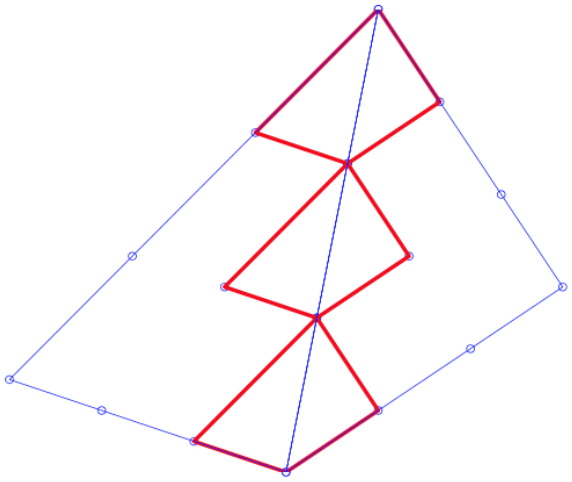

Figure 4: *Illustration of the constraints for C^1 smoothness, for a cubic on each triangle. Credit: de Boor (2009). A polynomial of an arbitrary degree on a triangle can be written be in what is called a B-form (as an expansion of what are called Bernstein basis polynomials), whose coefficients can be assigned to anchor points, drawn above as blue circles. The set of coefficients that are tied together in a linear equation needed for C^1 smoothness are highlighted as a red quadrilateral. We can see that the linear systems are entangled, since pairs of them overlap at an anchor point.*

polynomials in d dinensions, is

$$\sum_{\ell=0}^{k} \binom{d-1+\ell}{\ell} = \binom{d+k}{k}.$$

Now let's just think about cubic degree $k = 3$ in dimension $d = 2$, and try to understand the claim that it's tricky to construct a C^2 piecewise cubic, when we consider the "pieces" to be triangles. In fact, it's already going to be hard enough to construct a C^1 piecewise cubic. Suppose that we have two triangles sharing an edge e. Then on each triangle we have $\binom{5}{2} = 10$ parameters to define the cubic, and hence 20 parameters in total. Denoting by f, g the two cubics on our adjacent triangles, the C^1 condition says that

$$D^\alpha f(x) = D^\alpha g(x), \quad \text{for all } |\alpha| \leq 1 \text{ and all } x \in e.$$

It can be shown that this reduces to 10 constraints on the coefficients; but importantly, the structure of these constraints is such that, for a given set of values at the vertices of the triangles, it is unclear whether there will exist a set of coefficients that satisfies the constraints *and* meets the prescribed values at the vertices. This basically happens because the linear constraints are all entangled. Figure 4 gives an illustration.

To get around this problem we have to increase the degree of the polynomial. As it turns out, for quintic degree $k = 5$, this problem doesn't occur for two adjacent triangles in dimension $d = 2$, and we get what is sometimes called a "free" C^1 smoothness condition, where this kind of entanglement of constraints doesn't happen. One can ask: in dimension d, what is the requirement on the degree k (for a generic partition of points into simplices) such that we get a "free" C^s smoothness condition? The answer is, disappointingly,

$$k \geq s(d+1) + d.$$

We see that for "free" C^1 smoothness when $d = 2$, the minimal degree is indeed $k = 5$, and when $d = 3$ it is $k = 7$. This seems very wasteful.

This is not a complete show-stopper for multivariate splines and there is much more to the story than this but we'll not travel down this path. To learn more, two definitive references are de Boor et al. (1993); Lai and Schumaker (2007), and de Boor (2009) is a lighter, enjoyable review.

4.1 Tensor product splines

One way forward in the multivariate setting is just to take tensor products. Recall, a tensor product of two univariate functions f_1, f_2 on $[a, b]$ is denoted $f_1 \otimes f_2$, which is the function on $[a, b]^2$ defined as

$$(f_1 \otimes f_2)(x_1, x_2) = f_1(x_1) f_2(x_2).$$

Similarly, for any collection of $q \geq 2$ functions $f_1, \ldots, f_q$, we can define the tensor product $f_1 \otimes f_2 \otimes \cdots \otimes f_q$ as a function

$$(f_1 \otimes f_2 \otimes \cdots \otimes f_q)(x_1, x_2, \ldots, x_q) = f_1(x_1) f_2(x_2) \ldots f_q(x_q).$$

Thus given a k^{th} degree spline basis with r knots $t_1, \ldots, t_r$, which we denote by $g_1, \ldots, g_N$ for $N = r + k + 1$, we can take all d-way tensor products to form collection of functions on $[a, b]^d$:

$$g_{j_1} \otimes g_{j_2} \otimes \cdots \otimes g_{j_d}, \quad \text{for all } \ell = (j_1, j_2, \ldots, j_d) \in [N]^d,$$

where we use the abbreviation $[N] = \{1, \ldots, N\}$. By construction the linear span of this collection, i.e., functions of the form

$$f(x) = \sum_{j \in [N]^d} \beta_j \, g_{j_1}(x_1) \otimes \cdots \otimes g_{j_d}(x_d),$$

for coefficients β_j, $j \in [N]^d$, generates the space of tensor product splines

$$\Big\{ f_1 \otimes \cdots \otimes f_d : \text{each } f_\ell \text{ is a } k^{\text{th}} \text{ degree spline with knots in } t_1, \ldots, t_r, \text{ for } \ell = 1, \ldots, d \Big\}.$$

Figure 5 displays a few bivariate tensor products of B-splines.

To fit a tensor product spline from data, the simplest thing to do would be the *tensor product regression spline*, whose coefficients are given by solving

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^n \bigg(y_i - \sum_{j \in [N]^d} \beta_j \, g_{j_1}(x_{i1}) \otimes \cdots \otimes g_{j_d}(x_{id}) \bigg)^2.$$

Denoting by $A \otimes B$ the Kronecker product of matrices A, B, and denoting by $G_\ell \in \mathbb{R}^{N \times N}$ the basis matrix for dimension ℓ, with entries $[G_\ell]_{pq} = g_q(x_{p\ell})$, we can express the above in more compact form as

$$\underset{\beta}{\text{minimize}} \; \big\| Y - (G_1 \otimes \cdots G_d)\beta \big\|_2^2.$$

As usual, given the solution $\hat{\beta}$, we make predictions according to $\hat{f}(x) = \sum_{j \in [N]^d} \hat{\beta}_j \, g_{j_1}(x_1) \otimes \cdots \otimes g_{j_d}(x_d)$, and since $\hat{\beta}$ is linear in Y, the tensor product spline $\hat{f}$ is a linear smoother.

Few notes: there is no reason that the knots have to be the same along each dimension, that is only done here for simplicity. We could have also used regularization, though for "variationally-motivated" regularization we'd turn to thin plate splines (covered next).

4.1.1 What are these functions?

We should be clear that tensor product construction *does not "really" give us a multivariate spline*, if we maintain that a spline is a k^{th} piecewise polynomial that is also C^{k-1}. Why? Note that a tensor product of univariate k^{th} degree splines produces a function that is indeed C^{k-1}, and indeed a piecewise polynomial: it is a polynomial on each hypercube of the form

$$[t_{j_1}, t_{j_1+1}] \times \cdots \times [t_{j_d}, t_{j_d+1}].$$

But it is *not* necessarily a k^{th} degree polynomial on each hypercube; it is a tensor product of univariate k^{th} degree polynomials. Technically, out of this we can get a polynomial of degree up to k^d, and at the

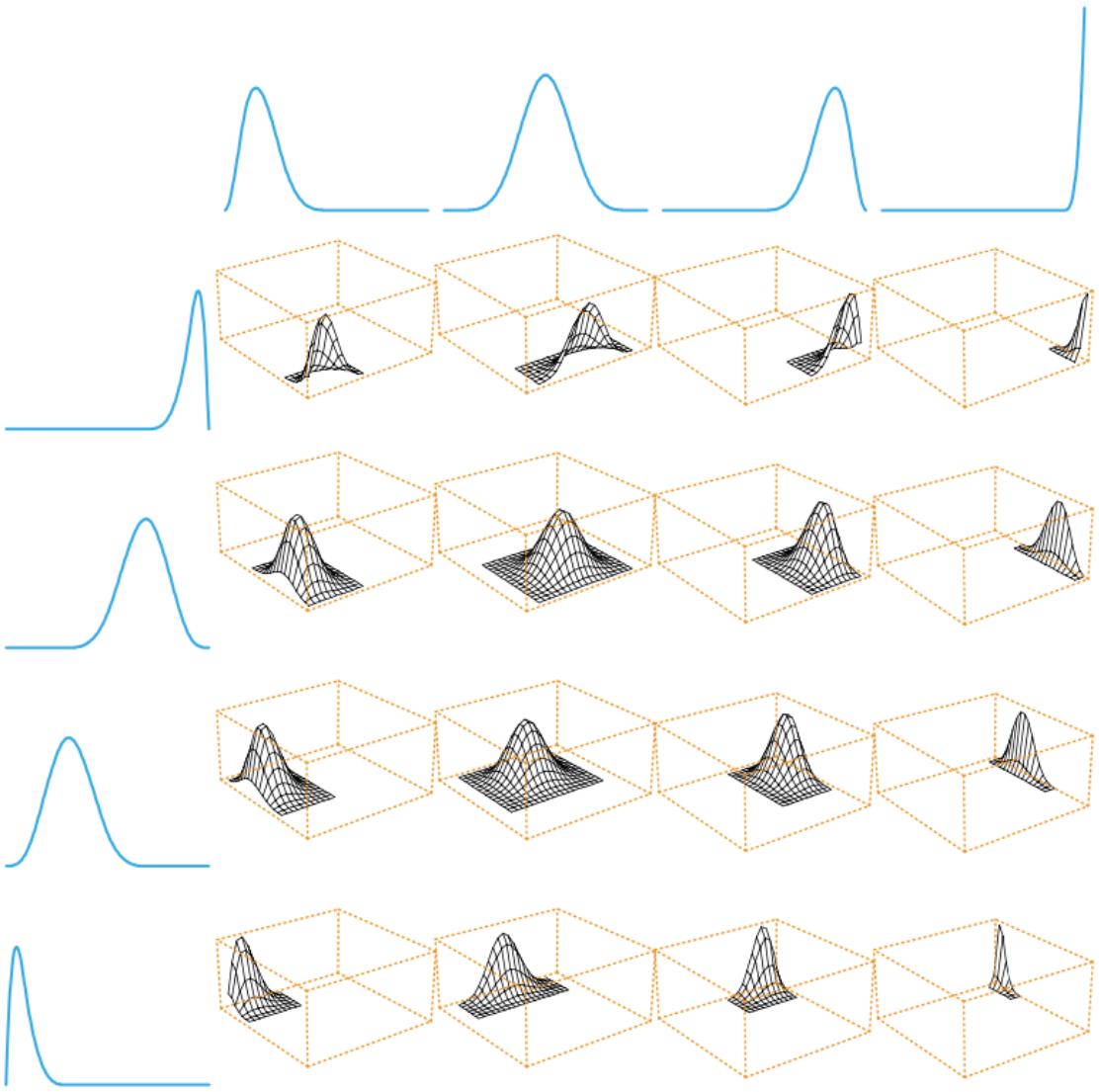

Figure 5: *Bivariate tensor products of cubic B-splines. Credit: Chapter 5.7 of Hastie et al. (2009).*

same time, we can't actually generate any polynomial of degree k^d. We are restricted to functions of the form

$$f(x) = \sum_{\alpha_1,\ldots,\alpha_d \leq k} \beta_\alpha \, x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}.$$

There is important difference here, compared to (8). As a concrete example, take $k = 1$ and $d = 2$: then the tensor product of univariate linear functions give us (what are called bilinear) functions of the form

$$1 + ax_1 + bx_2 + cx_1x_2,$$

which is *not* linear, because of the cross term x_1x_2. At the same time, we can't get any quadratic out of this, because we're missing the terms x_1^2 and x_2^2.

This seems to bothers some people (researchers, practitioners), but doesn't bother others—it really must depend on the problem setting or application one has in mind. Sometimes the highly anisotropic (coordinate-aligned) nature of a tensor product spline may be desirable, and other times it may not.

4.2 Thin plate splines

An alternative route is to try to extend the variational problem (2) that defined the univariate smoothing spline. At the outset, let's aim for the most general extension, which is to define an estimator by solving

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_U \sum_{|\alpha|=m} [D^\alpha f(x)]^2 \, dx, \tag{9}$$

for an integer order $m \geq 1$. Here U is a subset of $\mathbb{R}^d$ that contains the input points x_i, $i = 1, \ldots, n$, and the minimization is over all functions f for the which the criterion is well-defined and finite. For example, we can consider the minimization to be over the Sobolev space $W^{m,2}(U)$, which recall contains m times weakly differentiable functions with α^{th} derivative in $L^2(U)$, for $|\alpha| \leq m$. (There is a subtlety here about what we actually mean here for unbounded domain like $U = \mathbb{R}^d$. The problem is that the Sobolev space $W^{m,2}(\mathbb{R}^d)$ doesn't admit polynomials of degree $m-1$, whereas we'd like all of these to be allowed in the minimization. The trick is essentially to "quotient out" polynomials, which must be done with care; if you are curious to see the details, you can refer to Meinguet (1979); Wahba and Wendelberger (1980).)

Is this a good idea? It depends. When $2m > d$, problem (9) is well-defined in the sense that it admits a solution. This solution is what we'll eventually call the thin plate spline estimator (in a specific case). But when $2m \leq d$, problem (9) is *fundamentally ill-defined and lacks a solution altogether*, as explained next.

4.2.1 Return of the Sobolev embedding theorem

Let's recall what we learned from the Sobolev embedding theorem earlier (if you skipped it, since we did say it it was optional, then also that's fine—the key conclusion itself should still parse). When $2m > d$, which we called the supercritical regime, the Sobolev space $W^{m,2}(U)$ embeds continuously into a Hölder space $C^{r+\gamma}(U)$, so in particular it embeds continuously into $C^0(U)$, the space of continuous functions on U equipped with the L^∞ norm (this is the sup norm, and *not* the essential sup norm). Why is this important? Well, given a sequence such that $f_N \to f$ as $N \to \infty$ in $\|\cdot\|_{W^{m,2}(U)}$ norm, the continuous embedding property implies $f_N \to f$ as $N \to \infty$ in L^∞ norm, which implies $f_N(x) \to f(x)$ for all x. Thus we have shown a critical fact:

For $2m > d$, the point evaluation operator is continuous on $W^{m,2}(U)$.

This means we can do things like solve the variational problem (9).

But when $2m \leq d$, we do not get an embedding into $C^0(U)$, and it turns out that:

For $2m \leq d$, the point evaluation operator is not continuous on $W^{m,2}(U)$.

This means that there can be a sequence such that $f_N \to f$ as $N \to \infty$ in $\|\cdot\|_{W^{m,2}(U)}$ norm but $f_N(x) \not\to f(x)$ for some x. In fact, it's not hard to see this from first principles when $2m < d$. Let f be any infinitely differentiable "bump" function which is unimodal with a unique max at the origin, of $f(0) = 1$, and which is zero outside of the unit ℓ_2 ball $\{x : \|x\|_2 \leq 1\} \subseteq U$. Define $f_N(x) = f(Nx)$, for $N = 1, 2, 3, \ldots$, so that f_N collapses smoothly to a spike at the origin as $N \to \infty$. Then for any α,

$$\int [D^\alpha f_N(x)]^2 \, dx = N^{2|\alpha|-d} \int [D^\alpha f(u)]^2 \, du,$$

simply by a change of variables $u = Nx$. If $d > 2m$, then for any $|\alpha| \leq m$, we will have $N^{2|\alpha|-d} \leq N^{2m-d} \to 0$ as $N \to \infty$. In other words, $f_N \to 0$ (the zero function) as $N \to \infty$ in $\|\cdot\|_{W^{m,2}(U)}$ norm. However, recall $f_N(0) = 1$ for all N, so clearly we do not get pointwise convergence at the origin.

When $2m = d$, the construction needs to be adapted (we need to use a "multiresolution" average of bump functions, rather than a single bump function), but we can still show that point evaluation is not continuous by exhibiting a sequence for which $f_N \to 0$ in the $\|\cdot\|_{W^{m,2}(U)}$ norm, but $f_N(0) = 1$ for all N.

The lack of continuous point evaluation operator when $2m \leq d$ means that we *cannot* generally solve the variational problem (9), because its criterion (which features point evaluation) is not a continuous function of the optimization variable f.

4.2.2 Restrictions to the rescue

To circumvent the above issue, we can just restrict $2m > d$. Note that this forces us to take more derivatives as d grows, which is not really desirable (more on this later). In any case, when $d = 2$, it is valid to take $m = 2$, which yields the *thin plate spline* estimator, defined by solving

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_U \|\nabla^2 f(x)\|_F^2 \, dx, \tag{10}$$

where $\nabla^2 f$ denotes the weak Hessian (matrix of weak second derivatives) of f, and $\|\cdot\|_F$ is the Frobenius norm. This can be equivalently written as

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_U \left[\left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2} \right)^2 \right] dx. \tag{11}$$

Like the cubic smoothing spline problem (3), the thin plate spline problem (11) (or equivalently, (10)) has a representer theorem for its solution. Unlike the smoothing spline, the solution actually depends on the choice of the domain U. For $U = \mathbb{R}^d$, one can show that it suffices to consider functions of the form

$$f(x) = a + b^T x + \sum_{j=1}^{n} \beta_j \eta(\|x - x_j\|_2), \tag{12}$$

where η is the so-called *radial basis function*

$$\eta(r) = \frac{1}{16\pi} r^2 \log r^2,$$

(As $\log 0$ is undefined, we adopt a continuous extention at zero, and set $\eta(0) = 0$.) In order for the penalty in the problem (11) to be finite, a necessary and sufficient condition on the coefficients is that

$$0 = \sum_{i=1}^{n} \beta_j = \sum_{j=1}^{n} \beta_j x_j.$$

Note that there are 3 linear constraints here (since each x_i is in $\mathbb{R}^2$), and there are $n + 3$ parameters in (12), which means that in total, subject the constraints, we have n free parameters.

Using this constrained parametrization, we can translate (11) into a finite-dimensional problem. One can show that it is a constrained generalized ridge regression, thus its solution, the fitted coefficients $\hat{a}, \hat{b}, \hat{\beta}$, are linear in Y. As we make predictions according to $\hat{f}(x) = \hat{a} + \hat{b}^T x + \sum_{j=1}^{n} \hat{\beta}_j \eta(\|x - x_i\|_2)$, we see that the thin plate spline is a linear smoother. The story for what happens on a bounded domain U is similar, but a bit more complex. See Chapters 7.4–7.7 of Green and Silverman (1993) for precise details.

The function f in (12) often called a *polyharmonic spline*. Again, it is *not "really" a multivariate spline*, because it is not a piecewise polynomial. But let's emphasize an important property: η is symmetric around the origin. In fact, if we just look back at the criterion (10), we see that if we rotated the coordinate system by an orthogonal transform $V \in \mathbb{R}^{2\times 2}$, and take our new coordinates to be $z = Vx$, then we get

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(z_i))^2 + \lambda \int \underbrace{\|V \nabla^2 f(z) V^T\|_F^2}_{= \|\nabla^2 f(z)\|_F^2} \, dz,$$

where we used the fact that $\|VAV^T\|_F = \|A\|_F$ for any matrix A due to orthogonality of V. This means that the thin plate spline estimator is *rotationally invariant*: if we rotated the coordinate system of the input points, then the new solution would just be a rotation of the old solution. This is *not* true of tensor product splines.

What about more than 2 dimensions? When $d = 3$, the choice $m = 2$ is still valid, and the corresponding estimator is often still called the thin plate spline. The details are similar to the above. In fact, whenever

$2m > d$, the variational problem (9) has a representer theorem and admits a finite-dimensional solution which is a polyharmonic spline, and the problem reduces to a generalized ridge regression. See Chapter 7.9 of Green and Silverman (1993). Instead of pursuing these details, we'll simply move on to RKHS theory, as this will provide us a more general perspective.

But before moving on to RKHS theory, it's worth emphasizing restrictive the condition $2m > d$ is. Though it's a very simple calculation, Table 1 lists the requirement on m for $d = 1$ through 5. This restriction is obviously not desirable: the choice of smoothness order m (which then dictates the degree of the polyharmonic spline) should be up to the modeler, and not dictated by the ambient dimension.

Dimension	Restriction
$d = 1$	$m \geq 1$
$d = 2$	$m \geq 2$
$d = 3$	$m \geq 2$
$d = 4$	$m \geq 3$
$d = 5$	$m \geq 3$

Table 1: *Translating the supercritical condition $2m > d$, for each $d \leq 5$.*

4.2.3 Can we do anything else?

Yes. There is another way—we can "discretize" the variational problem (9), specifically, we can substitute the penalty with a discrete approximation based on a graph Laplacian, where the graph is built using the input points $x_1, \ldots, x_n$. This has the advantage of being (i) always well-defined, and (ii) more computationally efficient than solving (9) even when the latter is well-posed. In fact, beyond just being well-defined or computationally efficient, the graph-based approach can even be minimax rate optimal (over norm balls in Sobolev spaces) in regimes where the variational problem (9) is ill-posed. See Green et al. (2021), and also Green et al. (2023) for a related idea with even stronger theory. Of course, we're skirting a lot of details, but the takeaway you should know is this:

> *Consistent estimation over $W^{m,2}([0,1]^d)$—even rate optimal estimation—is still possible for general m, d. We can't use always use variational methods like (9), but we can do other things.*

However, to qualify the above statement, we *must* be in the Random-X setting, where x_i, $i = 1, \ldots, n$ are random. If the input points were fixed, then when $2m \leq d$, a construction using "bump" functions (just like the one we gave above) will show that consistent estimation is impossible.

5 RKHS methods

Lastly, we turn to RKHS methods. This will require covering some theory first in order to properly understand what's going on, but RKHS theory is arguably easier than Sobolev theory. Two great references on the former are Scholkopf and Smola (2002); Christmann and Steinwart (2008).

A **warning** before we begin: don't confuse any of what follows with kernel smoothing (as covered in the last lecture) or the kernel functions used therein! This is a completely separate concept.

5.1 Kernels, Hilbert spaces, RKHS

We start off by defining the ingredients needed for an RKHS. Throughout, let $\mathcal{H}$ be a vector space over $\mathbb{R}$. For our purposes: think of this as a space of real-valued functions. Recall, an *inner product* on $\mathcal{H}$ is a function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfying, for all $f, g, h \in \mathcal{H}$ and all $a, b \in \mathbb{R}$, the following three properties:

1. $\langle af + bg, h \rangle_{\mathcal{H}} = a\langle f, h \rangle_{\mathcal{H}} + b\langle g, h \rangle_{\mathcal{H}}$;
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$;
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ with equality iff $f = 0$.

We can always define a norm based on this inner product, denoted $\|\cdot\|_{\mathcal{H}}$, by

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}.$$

Equipped with its inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we call $\mathcal{H}$ an *inner product space*.

Hilbert spaces. An inner product space $\mathcal{H}$ is called a *Hilbert space* if it is complete. Recall that completeness means that Cauchy sequences converge to limits in the space. (Concretely, given any sequence f_n, $n = 1, 2, 3, \ldots$ in $\mathcal{H}$ such that for any $\epsilon > 0$, we have $\|f_n - f_m\|_{\mathcal{H}} < \epsilon$ for large enough n, m, it must hold that $\|f_n - f\|_{\mathcal{H}} \to 0$ as $n \to \infty$ for some $f \in \mathcal{H}$.)

Kernels. There are different ways to define kernels, and in fact different ways to arrive at an RKHS. One way is this. A *kernel* is function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a map $\phi : \mathcal{X} \to \mathcal{H}$ (for our purposes: think of this as a feature map) such that for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

Note that k is always symmetric and nonnegative, because $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is an inner product. Furthermore, it is easy to check from first principles that k satisfies the following property: for any $n \geq 1$ and $x_1, \ldots, x_n \in \mathcal{X}$, if we define a matrix $K \in \mathbb{R}^{n \times n}$ to have entries $K_{ij} = k(x_i, x_j)$, then

$$a^T K a \geq 0, \quad \text{for all } a \in \mathbb{R}^n.$$

A function k that is symmetric and satisfies the above property is called *positive semidefinite*.

Thus, from what we have just argued, a kernel function is always positive semidefinite. Remarkably, the converse is also true: any positive semidefinite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel. This means that *in order to define a kernel k, we don't have to explicitly specify the feature map ϕ; we just need to satisfy positive semidefiniteness*. Examples of kernels on $\mathcal{X} = \mathbb{R}^d$ include:

$$\begin{aligned}
&\text{polynomial kernel :} && k(x, y) = (1 + x^T y)^m, \\
&\text{exponential kernel :} && k(x, y) = \exp(x^T y), \\
&\text{Gaussian kernel :} && k(x, y) = \exp(-\|x - y\|_2^2 / \sigma^2).
\end{aligned}$$

Importantly, kernels need not be restricted to $\mathcal{X} \subseteq \mathbb{R}^d$; they can be defined over much more exotic spaces $\mathcal{X}$, like spaces of images or text strings, we just need a suitable (positive semidefinite) notion of similarity between objects.

Reproducing kernel Hilbert spaces. A Hilbert space $\mathcal{H}$ of functions is called a *reproducing kernel Hilbert space* (RKHS), with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, provided the following two conditions are satisfied:

1. for any $x \in \mathcal{X}$, the function $k(\cdot, x)$ is an element of $\mathcal{H}$;
2. for any function $f \in \mathcal{H}$ and $x \in \mathcal{X}$, it holds that $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

The functions $k(\cdot, x)$ in the first property are often called the "reproducers of evaluation". Accordingly, the second property is often called the "reproducing property". In particular, note that it implies

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}},$$

for any x, y. This means that the map $\phi(x) = k(\cdot, x)$ is a valid feature map for k.

Equivalent formulation of an RKHS. Here is an equivalent formulation of an RKHS that reveals a nice connection to Sobolev spaces. For $x \in \mathcal{X}$, denote by δ_x the corresponding evaluation operator on $\mathcal{H}$, which acts according to $\delta_x(f) = f(x)$. Then a Hilbert space of functions $\mathcal{H}$ is an RKHS if and only if all of the evaluation operators are continuous.

Recall that the Sobolev space $W^{m,2}(U)$, where $U \subseteq \mathbb{R}^d$, is a Hilbert space, for any $m \geq 0$. Recall also that point evaluation in a Sobolev space is continuous if and only if $2m > d$. Therefore, we learn the following: *a Sobolev space $W^{m,2}(U)$ is an RKHS if and only if $2m > d$.* This is an amazing fact.

A natural follow-up question: what is the kernel k corresponding to the Sobolev space, for $2m > d$? Well, it's a bit complicated. Are you ready for it? For $U = \mathbb{R}^d$, it's

$$k(x,y) = \int \frac{\exp(2\pi i(x-y)^\mathsf{T} u)}{1 + \sum_{0<|\alpha|\leq m} \prod_{j=1}^d (2\pi u_j)^{2\alpha_j}} \, du.$$

(The integral here—thus the kernel—only takes real values, since $\exp(2\pi i(x-y)^\mathsf{T} u) = \cos(2\pi(x-y)^\mathsf{T} u) + i(2\pi i(x-y)^\mathsf{T} u)$ and the integral of the imaginary part is zero because the corresponding integrand is odd.) While this certainly looks complicated, it simplifies in a few important cases. For $d = 1$ and $m = 2$, we get the "smoothing spline kernel":

$$k(x,y) = \frac{\sqrt{3}}{3} \exp\left(-\frac{\sqrt{3}|x-y|}{2}\right) \sin\left(\frac{|x-y|}{2} + \frac{\pi}{6}\right). \tag{13}$$

Meanwhile, for $d = 2$ and $m = 2$, we get the "thin plate spline kernel":

$$k(x,y) = \frac{1}{16\pi} \|x-y\|_2^2 \log \|x-y\|_2^2. \tag{14}$$

5.2 RKHS regression

After all that build up, this next part is going to be (surprisingly?) simple. Let $\mathcal{H}$ be an RKHS, and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be its kernel. We can now define an *RKHS regression estimator* by solving

$$\underset{f}{\text{minimize}} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{15}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm associated with $\mathcal{H}$. Note: here the input points satisfy $x_i \in \mathcal{X}$, and $\mathcal{X}$ can be completely arbitrary (it doesn't need to be a subset of $\mathbb{R}^d$).

Popular choices of k, as we already mentioned, include the polynomial kernel, Gaussian kernel, and exponential kernel, which gives novel nonparametric regression estimators—novel in the sense that you haven't seem them thus far in these lectures on nonparametric regresion. We can also choose particular kernels, as given in (13) and (14), to "reproduce" (no pun in intended) smoothing splines and thin plate splines.

5.3 Representer theorem

One of the main points of appeal of RKHS regression is the fact that we have a *representer theorem* for the infinite-dimensional problem (15). This effectively generalizes the results stated thus far on splines. The solution in (15) is unique and achieved at a function of the form

$$f(x) = \sum_{i=1}^n \beta_i k(x, x_i).$$

Note that this is a linear combination of the representers of evaluation. The proof of this result is actually not very hard. We'll walk through it on the homework.

5.4 Finite-dimensional form

From the representer result, we can rewrite (15) as

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^n \left(y_i - \sum_{j=1}^n \beta_j k(x_i, x_j)\right)^2 + \lambda \sum_{i,j=1}^n \beta_i \beta_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}}. \tag{16}$$

Using the reproducing property $\langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}} = k(x_i, x_j)$, and denoting by $K \in \mathbb{R}^{n\times n}$ the kernel matrix with entries $K_{ij} = k(x_i, x_j)$, problem (16) simplifies even further to

$$\underset{\beta}{\text{minimize}} \ \|Y - K\beta\|_2^2 + \lambda\beta^\mathsf{T} K\beta. \tag{17}$$

This is (once again) a generalized ridge regression. The solution in (17) is

$$\hat{\beta} = (K^\mathsf{T}K + \lambda K)^{-1}K^\mathsf{T}Y = (K + \lambda I)^{-1}Y,$$

where we have used the fact that K is symmetric and invertible. The RKHS estimator makes predictions according to $\hat{f}(x) = \sum_{j=1}^d \hat{\beta}_j k(x, x_j)$, and since $\hat{\beta}$ is linear in Y, it is (once again) a linear smoother.

References

Adreas Christmann and Ingo Steinwart. *Support Vector Machines*. Springer, 2008.

Carl de Boor. *A Practical Guide to Splines*. Springer, 1978.

Carl de Boor. The way things were in multivariate splines: A personal view. In *Multiscale, Nonlinear and Adaptive Approximation*, 2009.

Carl de Boor, Klaus Höllig, and Sherman Riemenschneider. *Box Splines*. Springer, 1993.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society: Graduate Studies in Mathematics, second edition, 2010.

Alden Green, Sivaraman Balakrishnan, and Ryan J. Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian regularization on neighborhood graphs. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021.

Alden Green, Sivaraman Balakrishnan, and Ryan J. Tibshirani. Minimax optimal regression over Sobolev spaces via Laplacian eigenmaps on neighborhood graphs. *Information and Inference*, 12(13):2423–2502, 2023.

Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall/CRC Press, 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, second edition, 2009.

Ming-Jun Lai and Larry L. Schumaker. *Spline Functions on Triangulations*. Cambridge University Press, 2007.

Jean Meinguet. Multivariate interpolation at arbitrary points made simple. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 30(2):292–304, 1979.

Bernhard Scholkopf and Alexander Smola. *Learning with Kernels*. The MIT Press, 2002.

Larry L. Schumaker. *Spline Functions: Basic Theory*. Cambridge University Press, third edition, 2007.

Bernard W. Silverman. Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, 12(3): 898–916, 1984.

Grace Wahba and James Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, 108(8):1122–1143, 1980.