Extensions of Thin-Plates Splines

Advanced Topics in Statistical Learning, Spring 2023

Zachary McNulty

May 4th, 2023

1 Review

As usual we consider the standard nonparametric regression setting. Specifically, given a regression function $f_0 : \mathbb{R}^d \to \mathbb{R}$ and samples $\{(x_i, Y_i)\}_{i=1}^n$ from:

$$Y_i = f_0(x_i) + \epsilon_i \quad \mathbb{E}\epsilon_i = 0,\ \sigma^2 := \mathbb{E}\epsilon_i^2 < \infty \tag{1}$$

for fixed covariates x_i (or equivalently, after conditioning upon the covariates), the goal was to find a good estimator $\hat{f}$ of f_0. Namely, we wanted to find an $\hat{f}$ that:

(i) Fit the training data well: $\sum_{i=1}^n (Y_i - \hat{f}(x_i))^2$ is small.

(ii) Is "smooth" so as to reduce the variance of our estimator.

This motivated the following variational problem:

$$\hat{f} = \underset{f \in W}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \sum_{|\alpha|=m} \binom{m}{\alpha} \int_{R^d} |D^\alpha f|^2 \, dx \tag{2}$$

where the sum is over all multi-indices α of degree m, the minimization is over the Sobolev space $W := W^{2,m}(\mathbb{R}^d)$ with smoothness parameter m (the largest class of functions for which the objective is well-defined), and $\lambda > 0$ is some regularization hyperparameter. Intuitively, the first term in (2) penalizes a poor fit while the second term penalizes a lack of smoothness, and λ determines the trade-off between these two aims.

We saw that, as a consequence of the Sobolev embedding theorem, the variational problem (2) had a solution only when $2m > d$ and was ill-posed otherwise. From now on, we consider only the $2m > d$ case. For $d = 1$ we obtained a true spline, the smoothing spline, but for $d = m = 2$ the solution is of the form:

$$\hat{f}(x) = a + b^T x + \sum_{j=1}^n \beta_j \eta(||x - x_j||_2) \tag{3}$$

where $\eta(r) = -\frac{1}{16\pi} r^2 \ln(r^2)$ is a radial basis function[1] and the coefficients are subject to the constraints $\sum \beta_j = \sum \beta_j x_j = 0$. Note that the solution (3) consists of two parts: a linear part $a + b^T x$ and a radial part $\eta(||x - x_j||_2)$. The former is natural because any linear term has zero roughness penalty as it vanishes under the second derivative while the latter is natural because both it and the roughness penalty are invariant under rotations of the coordinate system. Again the benefit of this is that it reduced the intractable infinite dimensional optimization problem (2) into a finite-dimensional regression problem on the coefficients (a, b, β_j).

[1] $\eta(||x - x_0||_2)$ is a fundamental solution to the biharmonic equations $\Delta^2 f = 0$.

2 Exploring New Loss Functions

The squared loss $\rho(z) := z^2$ on the residuals in (2) is nice for many reasons: its simplicity makes it very theoretically tractable, it is computationally efficient, and it has been studied in many settings within the statistical literature. Nonetheless, it has some major drawbacks. For one, it is well-known such a loss is sensitive to outliers. This will be especially problematic if the errors ϵ_i have heavier tails than the traditional Gaussian tails for $\epsilon_i \sim N(0, \sigma^2)$. Thus it may be preferable to have a loss function that increases more gradually as $|x| \to \infty$, especially if we wish to consider different error models. Secondly, squared loss is symmetric. In most practical applications it is often more costly to overestimate than underestimate (or vice versa), necessitating an asymmetric loss function. Hence it is worthwhile extending our thin plate splines to other loss functions. This motivates the new variational problem where we replace the squared loss by a general loss function $\rho : \mathbb{R} \to \mathbb{R}_{\geq 0}$:

$$\hat{f} = \underset{f \in W}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - f(x_i)) + \lambda \sum_{|\alpha|=m} \binom{m}{\alpha} \int_{R^d} |D^\alpha f|^2 \, dx. \tag{4}$$

For compactness of notation, we will denote the objective of (4) as $L_n(f)$ and the smoothness penalty term as $I_m^2(f)$. Our analysis of this problem below follows the work of Kalogridis (2022) which calls the solutions to (4) ***M-type thin plate splines*** (MTPS). Of course, to hope to say anything in general about problem (4) we must make some regularity assumptions on the loss function.

Assumptions

(A1) ρ is convex and Lipschitz.

(A2) There exists a constant $\kappa > 0$ so that if $|t| \leq \kappa$:

$$\inf_n \min_{i \leq n} \mathbb{E}_\epsilon \left[\rho(t + \epsilon_i) - \rho(\epsilon_i) \right] \geq \kappa t^2$$

(A3) The covariates $\{x_i\}_{i=1}^{\infty}$ are contained in a "nice" set.

Under these assumptions, we will outline the work of Kalogridis (2022) which develops a rate of convergence for the estimator (4) and shows that such an estimator must take a form analogous to one of (3). Let's first spend a moment trying to understand these assumptions.

Convexity is a natural assumption on the loss function because it makes (4) a convex optimization problem, and hence will at least be somewhat tractable to solve. Moreover, we will see it guarantees us the existence of a solution to (4). The Lipschitz assumption is mostly just to aid in the proof, but is not entirely harmless either: for one, squared error loss is not Lipschutz. Nonetheless, many other loss functions satisfy this property, and it is natural to want some level of continuity in a loss function.

Assumption (A2) is saying we want sufficiently fast growth of the loss function near zero. For the squared error loss $\rho(x) = x^2$ this condition is naturally satisfied:

$$\mathbb{E}_\epsilon((t + \epsilon_i)^2 - \epsilon_i^2) = \sigma^2 t^2$$

Hence (A2) is essentially saying the function grows quadratically (or faster) at least locally around zero. This will be important in proving the rate of convergence of the MTPS to the true f_0. If ρ grows too slowly near zero it will be difficult for the goodness of fit measure $n^{-1} \sum \rho(Y_i - f(x_i))$ to differentiate between f_0 and other functions f which lie close to f_0 at the covariates x_i. Hence the roughness penalty may begin to dominate too much, interfering with the desired convergence.

Lastly, (A3) just rules of degenerate choices of the covariates x_i by assuming some level of uniformity and regularity of their spacings. Essentially, it says that any newly sampled covariate x_0 will be close to one of the old covariates most of the time. As our functions are pretty smooth, this the performance at the new point will not be too much worse than at the covariates x_i. This will be essential for our development of the rate of convergence as it allows us to pass from the simpler setting of studying $||\hat{f} - f_0||_{L^2(\mathbb{P}_n)}$ to $||\hat{f} - f_0||_{L^2(\mathbb{P})}$ which might be of more interest.

2.1 Existence of Solutions

In the original thin plate spline (2) one obstacle was that the penalty term $I_m(f)$ has a nontrivial null space: $I_m(p) = 0$ for any polynomial $p : \mathbb{R}^d \to \mathbb{R}$ with total degree less than m. This can be problematic because we can replace any $f \in W$ by $f + p$ without changing our roughness penalty. Namely, we can always "wiggle" our fit to the data by a polynomial with no extra cost in the penalty.

This could be a problem when it comes to existence of a solution for two reasons. Firstly, if we could continuously tweak our function f by a series of polynomials to get a better and better fit, we may never achieve the infimum in (4). Similarly, given two different functions f, g it might be possible to find different polynomials p_1, p_2 to perturb these functions and make $f + p_1, g + p_2$ close. It turns out that these two problems are essentially the only things that can go wrong. Namely, given a basis $\phi_1, ..., \phi_M$ for the vector space polynomials of degree less than m, a space of dimension $M := \binom{m+d-1}{d}$, then as long as:

$$\min_{\alpha \in \mathbb{R}^M} \left|\left| Y - \sum_{j=1}^{M} \alpha_j \phi_j(x) \right|\right|_2$$

has a unique minimizer, we have a unique solution to (2). This is basically saying as long as the polynomial regression problem is well-posed, the infinite dimensional problem is as well. In the same vein, for general loss function ρ we obtain:

Theorem 1 (Existence of MTPS)**.** *If the optimization problem:*

$$\min_{\alpha \in \mathbb{R}^M} \sum_{i=1}^{n} \rho \left(Y_i - \sum_{j=1}^{M} \alpha_j \phi_j(x_i) \right)$$

has a unique solution, then (4) has a minimizer in W.

The justification for this will become clear in the following section. There, we will see the optimal solution will decompose into two parts: a polynomial fit (a fit over the nullspace) and an RKHS fit (a fit over everything else). The RKHS part always has a unique solution, so the only problems can occur in the polynomial fit.

2.2 Representer Theorem

Due to the infinite-dimensional nature of the optimization problem (4), the existence guarantee of theorem 1 is of little practical purpose. As with the standard TPS, our real saving grace will be the fact that we can show the solution must take a particular form, and this form can be found by regression over a finite-dimensional parameter space.

Theorem 2 (Representer Theorem)**.** *There exists a minimizer of $L_n(f)$ of the form:*

$$\hat{f}(x) = \sum_{i=1}^{n} \gamma_i \eta_{m,d}(||x - x_i||_2) + \sum_{j=1}^{M} \delta_j \phi_j(x) \tag{5}$$

where the coefficients are subject to the constraints:

$$\Phi^T \gamma := \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_M(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_M(x_2) \\ \vdots & & \ddots & \vdots \\ \phi_1(x_n) & \phi_2(x_n) & \dots & \phi_M(x_n) \end{bmatrix}^T \gamma = 0$$

Namely, analogous to (3), the solution must be the sum of two parts: a radial part and a polynomial. Here $\eta_{m,d}(||x - x_i||)$ are fundamental solutions to the m-iterated Laplacian Δ^m which take the form:

$$\eta_{m,d}(r) \propto \begin{cases} x^{2m-d}\log(x) & \text{d is even} \\ x^{2m-d} & \text{d is odd} \end{cases}$$

The condition on γ merely states γ is perpendicular to the space of all polynomial fits, namely vectors of the form $[p(x_1), ..., p(x_n)]$ for some polynomial p of degree less than m. The proof is very similar to our proof of the representer theorem for the smoothing spline. In particular, it can be shown that given any other $g \in W$ we can find an $\hat{f}$ of the form (5) which can interpolate $(x_i, g(x_i))$ and which is at least as smooth as g.

Let's try to make sense of where this representation is coming from. The inclusion of the polynomial is obvious: any polynomial of degree less than m has zero roughness penalty, and hence we can include it in our fit basically for free. To see where the $\eta_{m,d}$ are coming from observe that a few applications of integration by parts yields:

$$I_m^2(f) := \sum_{|\alpha|=m} \binom{m}{\alpha} \int_{\mathbb{R}^d} |D^\alpha f|^2 = \int_{\mathbb{R}^d} f\Delta^m f \tag{6}$$

Hence we can view the roughness penalty as the action of a seminorm induced by the semi-inner product:

$$\langle f, g\rangle = \sum_{|\alpha|=m} \binom{m}{\alpha} \int_{\mathbb{R}^d} D^\alpha f D^\alpha g = \int_{\mathbb{R}^d} f\Delta^m g$$

If g is a fundamental solution to Δ^m, namely if $\Delta^m g = \delta_x$ then we have $\langle f, g\rangle = \int f\delta_x = f(x)$. Hence $\langle \cdot, g\rangle$ produces exactly the "evaluation at x" linear functional. In particular, g is playing exactly the same role as the reproducing kernel $k(\cdot, x)$ in an RKHS. Recall from RKHS theory that given a RKHS $\mathcal{H}$ we have the analogous representation theorem for squared error loss:

$$\hat{f} = \underset{f\in\mathcal{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda||f||_{\mathcal{H}}^2 \to \hat{f} = \sum_{i=1}^n \beta_i k(\cdot, x_i) \tag{7}$$

Thus Theorem 2 is essentially saying our fit breaks down into two parts: a fit on the nullspace of our seminorm $I_m(f)$, namely a polynomial fit, and an RKHS fit on the orthogonal complement using the above (semi) inner product[2].

Since $\eta_{m,d}(||x - x_i||)$ is a fundamental solution to the iterated Laplacian Δ^m and $\Delta^m\phi = 0$ for any polynomial of degree less than m, using (6) we can see $\hat{f}$ of the form (5) has penalty:

$$\begin{aligned} I_m(\hat{f}) &= \int \left(\sum_{i=1}^n \gamma_i \eta_{m,d}(||x - x_i||_2) + \sum_{j=1}^M \delta_j\phi_j(x)\right) \cdot \left(\sum_{k=1}^n \gamma_k \delta_{x_k}\right) \\ &= \sum_{i,j=1}^n \gamma_i\gamma_j\eta_{m,d}(||x_j - x_i||_2) \end{aligned}$$

where the polynomial term vanishes due to the $\Phi^T\gamma = 0$ condition. Hence the optimization problem (4) becomes a form of generalized ridge regression where the standard squared-error loss has been replaced by our new convex loss ρ:

$$\lim_{\gamma\in\mathbb{R}^n, \delta\in\mathbb{R}^M} \left[\frac{1}{n}\sum_{i=1}^n \rho\left(Y_i - \hat{f}(x_i)\right) + \lambda\gamma^T\Omega\gamma\right] \quad s.t. \quad \Phi^T\gamma = 0$$

[2]There are some technical details I have left out on how this decomposition is done. More details can be found here or in section 2 of Beatson et al. (2018).

where $\Omega_{ij} = \eta_{m,d}(||x_i - x_j||_2)$. Hence our problem reduces to a generalized form of ridge regression. We can convert this to an unconstrained problem simply by replacing γ by $Q\beta$ where the columns of Q form a basis for $Null(\Phi^T)$. The above is a common convex optimization problem which can be easily solved by a variety of methods.

2.3 Rate of Convergence

Now that we know a solution exists, the next natural question is how well such a solution will approximate the true mean function f_0. We hope that as the amount of data we gather grows, our MTPS eventually gets close to f_0. In this section, we exactly quantify the rate of this convergence.

Theorem 3 (Rate of Convergence MTPS)**.** *If we choose penalty parameter $\lambda \sim n^{-\frac{2m}{2m+d}}$ then there exists a sequence $\hat{f}_n$ of MTPS which are minimizers to (4) and satisfy:*

$$||\hat{f}_n - f_0||_{L^2(\mathbb{P}_n)} = O_P(n^{-\frac{2m}{2m+d}}) \quad \textit{and} \quad I_m(\hat{f}_n) = O_P(1).$$

Moreover, the same rate is achieved under $L^2(\mathbb{P})$ under assumption (A3).

The proper choice of λ comes from attempting to balance the asymptotic variance and bias of our given estimators. Note that the rate gets better as we increase our smoothness parameter m, but this comes at a computational cost: the number of parameters in (5) grows with $M \sim (m+d)^d$. Moreover, recall our representer theorem says that the MTPS is essentially a polynomial of degree at most m plus some radial part. Hence as we increase m, in a sense we are allowing our estimator to be less smooth as we are allowing for a high degree polynomial to be included in the fit.

Furthermore, just as with kNN and kernel smoothing, this rate too suffers from the curse of dimensionality. Namely, if we desire an error of at most ϵ than by setting $n^{-2m/(2m+d)} = \epsilon$ we see we need a sample number on the order of $(1/\epsilon)^{1+d/2m}$, exponentially large in the desired error rate.

Let's outline a few ideas from the proof. This is where our assumptions (A2) and (A3) will start to come into play. Let $M_n(f) := n^{-1}\sum \rho(Y_i - f(x_i))$:

1. **Interpolation**: Given an optimal solution $\hat{f}_n$ to (4) we define $\tilde{f} := \alpha\hat{f}_n + (1-\alpha)f_0$ for $\alpha \in (0,1)$ to be chosen later. Convexity of our optimization problem L_n and optimality of $\hat{f}_n$ imply $L_n(\tilde{f}) \leq L_n(f_0)$ no matter the choice of α.
2. **Localize around $\mathbf{f_0}$**: choose α small enough so $||\tilde{f} - f_0||_{L^2(\mathbb{P}_n)}$ is small. Note that by definition:

   $$||\tilde{f} - f_0||_{L^2(\mathbb{P}_n)} = \alpha||\hat{f}_n - f_0||_{L^2(\mathbb{P}_n)}$$

   Hence by controlling the left hand side, we also get bounds on the right which is what we desire. In particular, we will try to bound both sides of the following in terms of $||\tilde{f} - f_0||_{L^2(\mathbb{P}_n)}$:

   $$\begin{aligned} &L_n(\tilde{f}) + \mathbb{E}(M_n(\tilde{f}) - M_n(f_0)) \leq +L_n(f_0) + \mathbb{E}(M_n(\tilde{f}) - M_n(f_0)) \\ \rightarrow\ &\mathbb{E}(M_n(\tilde{f}) - M_n(f_0)) + I_m^2(\tilde{f}) \leq [M_n(f_0) - \mathbb{E}M_n(f_0)] - [M_n(\tilde{f}) - \mathbb{E}M_n(\tilde{f})] + I_m^2(f_0) \end{aligned}$$

   We use (A2) to lowerbound the left hand side. Namely because $\mathbb{E}Y_i = f_0(x_i)$ then $Y_i - f_0(x_i)$ is mostly close to zero. Hence if $||\tilde{f} - f_0||_{L^2(\mathbb{P}_n)}$ is large, due to the growth of ρ near zero this implies $M_n(\tilde{f}) - M_n(f_0)$ is likely large as well.
3. **Entropy**: For the upperbound we need to control the size of the set of "good estimators", those with good fit (in squared error sense) and good smoothness: $\{f \in W : ||f||_{L^2(\mathbb{P}_n)} \leq 1, I_m(f) \leq R\}$. We need this to get a better idea of what $M_n(\tilde{f})$ looks like.
4. **Approximation**: Here we use (A3) to extend from the $L^2(\mathbb{P}_n)$ to $L^2(\mathbb{P})$. Basically, as (A3) tells us a new point x_0 won't be too far away from our covariates x_i most of the time and as our functions are pretty smooth, we can discretize our space at the covariates x_i without much loss. Namely, $||\hat{f} - f_0||_{L^2(\mathbb{P})}$ is well-approximated by $||\hat{f} - f_0||_{L^2(\mathbb{P}_n)}$.

3 A New Domain: Splines on the Sphere

For some applications it is more natural to have a domain different from some open subspace of $\mathbb{R}^d$. For example, if the underlying function $f_0 : \mathbb{R} \to \mathbb{R}$ is periodic it is natural to impose a further restriction on the standard smoothing spline: it and its derivatives ought to be equal at the two endpoints. Since any periodic function $f_0 : \mathbb{R} \to \mathbb{R}$ can be viewed equivalently as a function $f_0 : S^1 \to \mathbb{R}$ defined on the circle, we can view this problem as trying to fit a spline on circle. More generally, we might be interested in fitting a thin plate spline on the hypersphere S^{d-1}. To give a concrete example, if one is trying to predict the value of some meteorological quantity across all of Earth, it is convenient to work on the domain S^2 to account for the fact that we expect continuity in these quantities as we circumnavigate the globe.

Golomb (1968) covers the S^1 case while Wahba (1981) extends this to S^2 and provides an explicit (but not closed form) solution for the associated reproducing kernels, and laid the foundation for much of the following work. More recent work by Beatson et al. (2018) gives the closed form expression in S^{d-1} via a recurrence for these kernels. In this section, we will outline some of this work.

Let's start with some motivation, starting on the circle $S^1 \subseteq \mathbb{R}^2$. We would like to find a way of making sense of our roughness penalty $I_2(f)$ on this new domain. Since I_2 is defined in terms of cartesian coordinates, our first idea might be to try to change to polar coordinates which are more natural for working on the circle. To make the outcome of this transformation more clear, let's manipulate I_2 a bit. As we have noted before, a few applications of integration by parts yields:

$$I_2^2(f) := \int_{\mathbb{R}^2} \left(f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2\right) = \int_{\mathbb{R}^2} f\Delta^2 f \; dx \; dy$$

where Δ is the Laplace operator. Rewriting Δ in polar coordinates gives:

$$\Delta f = f_{rr} + \frac{1}{r} f_r + \frac{1}{r^2} f_{\theta\theta}$$

If we want to consider functions f on S^1, then we can just think of choosing $r = 1$ and dropping the dependence of r (the f_r, f_{rr} terms) from the Laplacian. Hence our roughness penalty becomes $I_2(f) = \int_0^{2\pi} f(\theta) f^{(4)}(\theta) \; d\theta$. We can repeat the above argument for general m, d to get that:

$$I_m(f) = (-1)^m \int_{\mathbb{R}^d} f\Delta^m f \; dx_1 \; ... \; dx_d \tag{8}$$

and so understanding the roughness penalty on S^{d-1} just boils down to rewriting the Laplacian Δ in terms of the relevant spherical coordinates and dropping all dependencies on the radius r. This yields the ***Laplace-Beltrami operator*** Δ_* which for the sphere S^2 is of the form:

$$\Delta_* f = \frac{1}{\sin^2(\theta)} f_{\phi\phi} + \frac{1}{\sin(\theta)} \left(\sin(\theta) f\right)_\theta$$

Now that we have our roughness penalty sorted out, we can determine the space of functions we would like to optimize over to find our thin plate spline. Of course the natural choice is the largest space of functions for which the penalty (8) is well-defined. As we will see below, determining the proper class will rely heavily on the natural connection between the Fourier series and periodic functions as well as its higher dimensional analogous. We start by using our previous work as a motivating example for what is to come.

Recall that above we saw the solutions (2) were of the form:

$$\hat{f}(x) := p(x) + \sum \gamma_i \eta(||x - x_i||)$$

where p was a polynomial of degree less than m and $\eta(r)$ was a fundamental solution to the m-iterated Laplacian $\Delta^m u = \delta_0$. The former came as a basis for the nullspace of the standard iterated Laplacian Δ^m and the latter came from a fundamental solution to Δ^m and applying RKHS theory. In the same way, we

can develop a representer theorem for thin plate splines on the sphere by (i) determining the nullspace of Δ_*^m and (ii) using a fundamental solution to Δ_*^m to develop a RKHS.

To accomplish these two aims, we start just by trying to diagonalize our operator Δ_*. It turns out the **spherical harmonics** $\{Y_j^k\}$ for $k \geq 0, 1 \leq j \leq N_{d,k}$[3], which are the restrictions of the real harmonic homogeneous polynomials of degree k from $\mathbb{R}^d$ to S^{d-1}, are precisely our eigenfunctions with eigenvalues:

$$\Delta_* Y_j^k = -k(k+d-2)Y_j^k$$

Note that this implies the nullspace of Δ_*^m is precisely just the span of Y_0^1 which turns out to be the set of constant functions. Curiously, unlike before we no loneger have any dependence between the nullspace and the smoothing parameter m.

The particular form these spherical harmonics take is not super important, but these spherical harmonics have the crucial property that they form an orthonormal system for $L^2(S^{d-1})$ under the standard surface area measure. We can use this to decompose an arbitrary element $f \in L^2(S^{d-1})$ with respect to this system:

$$f = \sum_{k=0}^{\infty} \sum_{j=1}^{N_{d,k}} \left\langle f,\ Y_j^k \right\rangle_{L^2(S^{d-1})} Y_j^k$$

Parseval's identity and the fact the Y_j^k are eigenfunctions to Δ_* imply that:

$$||\Delta_*^m f||_2^2 = \sum_{k=1}^{\infty} (k(k+d-2))^m \sum_{j=1}^{N_{d,k}} \left\langle f, Y_j^k \right\rangle_{L^2}^2$$

Letting $\mathcal{F}_{d,m}$ be space of $f \in L^2(S^{d-1})$ where this is finite gives us a natural choice of domain for our thin plate splines. From here, we can repeat the ideas from Section 2.2. Namely we can define the natural semi-inner product on this space as:

$$\langle f, g \rangle_{d.m} := \sum_{k=1}^{\infty} (k(k+d-2))^m \sum_{j=1}^{N_{d,k}} \left\langle f, Y_j^k \right\rangle_{L^2} \cdot \left\langle g, Y_j^k \right\rangle_{L^2} \tag{9}$$

This becomes an inner product once we restrict to the space $\mathcal{F}_{d,m}^0$ of functions $f \in \mathcal{F}_{d,m}$ with $\left\langle f, Y_0^1 \right\rangle = 0$. Namely the subset of $\mathcal{F}_{d,m}$ orthogonal to the constant functions. Under this inner product Beatson et al. (2018) shows:

Theorem 4. *Under the inner product (9) the space $\mathcal{F}_{d,m}^0$ is an RKHS with kernel K function:*

$$K_{d,m}(x,y) := \sum_{k=1}^{\infty} (k(k+d-2))^{-m} N_{d,k} W_k^{\lambda}(x^T y)$$

*where $\lambda = \frac{d-2}{2}$ and W_n^{λ} is the **Gegenbauer polynomial** of order λ normalized so $W_k^{\lambda}(1) = 1$. Specifically:*

$$N_{d,k} W_k^{\lambda}(x^T y) = \sum_{j=1}^{N_{d,k}} Y_k^j(x) Y_k^j(y)$$

This kernel has feature map $\phi : S^{d-1} \to \mathcal{F}_{d.m}^0$ sending x to the function $f \in \mathcal{F}_{d,m}$ with Fourier coefficients:

$$\left\langle f, Y_j^k \right\rangle_{L^2} = \begin{cases} 0 & k = 0, j = 1 \\ (k(k+d-2))^{-m/2} Y_k^j(x) & otherwise \end{cases}$$

[3] $N_{d,k} = \binom{k+d-1}{k} - \binom{k+d-3}{k-2}$

A simple calculation shows this indeed has the reproducing property:

$$\langle f, K(\cdot,x)\rangle_{d,m} = \sum_{k=1}^{\infty}(k(k+d-2))^m \sum_{j=1}^{N_{d,k}} \left\langle f, Y_j^k \right\rangle_{L^2} \left\langle K(\cdot,x), Y_j^k \right\rangle_{L^2} = \sum_{k=1}^{\infty}\sum_{j=1}^{N_{d,k}} \left\langle f, Y_j^k \right\rangle_{L^2} Y_j^k(x) = f(x)$$

Moreover, just as was the case for $\eta_{m,d}(||x-x_i||)$ the kernel functions $K(\cdot,x)$ can be viewed as the fundamental solution to Δ_*^m. Using the fact Y_j^k are eigenfunctions of Δ_*^m:

$$\Delta_*^m K_{d,m}(\cdot,x) = \sum_{k=1}^{\infty}(k(k+d-2)^{-m} \sum_{j=1}^{N_{d,k}} Y_k^j(x)\Delta_*^m Y_k^j = \sum_{k=1}^{\infty}\sum_{j=1}^{N_{d,k}} Y_k^j(x) Y_k^j$$

Note that the right hand side is precisely the Fourier series of the function $f \in L^2(S^{d-1})$ satisfying $< f, Y_k^j >_{L^2} = Y_k^j(x)$. Namely, f is playing the role of the dirac function δ_x.

Hence as before we see our fit breaks into two parts: a fit over the constant functions and an RKHS fit over the orthogonal space $\mathcal{F}_{d,m}^0$ whose kernel function is essentially just the fundamental solution of the underlying iterated Laplacian Δ_*^m. Thus just as before we get a representer theorem

Theorem 5 (Representer Theorem for Splines on S^{d-1})**.** *The solution to variational problem:*

$$\underset{f\in\mathcal{F}_{d,m}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int_{S^{d-1}} f\Delta_*^m f$$

must take the form for some constant c:

$$\hat{f}(x) = c + \sum_{i=1}^{n}\beta_i K(x,x_i)$$

However, this is not as useful as it might seen because Theorem 4 only gives the kernel $K(\cdot,x_i)$ in the form of an infinite series which is not amenable to regression. Certainly we could get approximate kernel functions simply by taking the partial sums, but we would prefer a nice closed form expression like we had for $\eta_{m,d}$ in Theorem 2. The main contribution of Beatson et al. (2018) was the formulation of a recurrence relation for the associated function $\kappa : \mathbb{R} \to \mathbb{R}$ related to the kernel K by $K(x,y) = \kappa(x^T y)$. Since K only depends on x, y through $x^T y$ this is well-defined. This at least allows for explicit calculation of the kernel in cases of small m, d but does not give a general formula. But since most applications care simply about the $d = 3$ case, this is of course practically very useful.

3.1 Further Extensions

There is nothing special about the sphere S^{d-1} discussed in the previous section. Namely, it makes sense to define thin plate splines onto any domain in which one can make sense of derivatives, specifically the Laplacian. In particular, these ideas would all make sense on any Riemannian manifold, a space for which the Laplace-Beltrami operator is well-understood. Work by Steinke et al. (2008) explores some of these ideas and their applications to computer graphics.

Another interesting extension is to consider different penalty/energy functions. Namely, in essence all of the above work boils down to finding the appropriate RKHS associated to a given energy functional:

$$E(f) = \langle f, f\rangle = \sum_{|\alpha|=m} \binom{m}{\alpha} \int D^\alpha f \cdot D^\alpha f$$

Of course there are many other potential energy structures induced by an inner product structure like this one, and under the right choice of underlying space it is possible to find the inducing kernel. I would be curious about the interpretations of different differential operators in terms of desirable "smoothness" properties of the associated optimal solutions. In the literature, splines with penalty of the form $\int |Lf|^2$ for a linear operator L are called L-splines. See chapter 10 of Schumaker (2007) for more details.

References

Rick K. Beatson, , and Wolfgang zu Castell and. Thinplate splines on the sphere. *Symmetry, Integrability and Geometry: Methods and Applications*, aug 2018. doi: 10.3842/sigma.2018.083. URL https://doi.org/10.3842%2Fsigma.2018.083.

Michael Golomb. Approximation by periodic spline interpolants on uniform meshes. *Journal of Approximation Theory*, 1(1):26–65, 1968. ISSN 0021-9045. doi: https://doi.org/10.1016/0021-9045(68)90055-5. URL https://www.sciencedirect.com/science/article/pii/0021904568900555.

Ioannis Kalogridis. Robust thin-plate splines for multivariate spatial smoothing. 2022.

Larry Schumaker. *L-Splines*, page 420–461. Cambridge Mathematical Library. Cambridge University Press, 3 edition, 2007. doi: 10.1017/CBO9780511618994.012.

Florian Steinke, Matthias Hein, Jan Peters, and Bernhard Schölkopf. Manifold-valued thin-plate splines with applications in computer graphics. *Computer Graphics Forum*, 27(2):437–448, 2008. doi: https://doi.org/10.1111/j.1467-8659.2008.01141.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2008.01141.x.

Grace Wahba. Spline interpolation and smoothing on the sphere. *SIAM Journal on Scientific and Statistical Computing*, 2(1):5–16, 1981. doi: 10.1137/0902002. URL https://doi.org/10.1137/0902002.