5. Basis Expansions and Regularization

오영민

Index

Introduction

- It is extremely unlikely that the true function $f(X)$ is actually linear in X
- And representing $f(X)$ by a linear model is usually convenient.
- We discus popular methods for moving beyond linearity

Introduction

- Denote by $h_m(X) : \mathbb{R}^p \to \mathbb{R}$ the m-th transformation of X, $m = 1,...,M$
- We then the model $f(X) = \beta_1 h_1(X) + \ldots + \beta_M h_M(X),$ a linear basis expansion in X.
- The basis functions h_m have been determined, the models are linear in these new variables.

Regression Splines

Splines

- We assume that X is one-dimensional. A piecewise polynomial function $f(X)$ is obtained by dividing the domain of X into contiguous intervals, and representing X by a separate polynomial in each interval.
- A spline f of degree M with knots at $\xi_1 < \ldots < \xi_K$:
  - f is polynomial of degree M on each of $(-\infty, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_K, \infty)$
  - $f^{(l)}$ is continuous at each of $\xi_1, \ldots, \xi_K$, for all $l = 0,...,M-1$.

Regression Splines

Spline bases

- **Truncated power bases:**
  - $h_j(X) = X^{j-1},\ j = 1,\ldots,M+1$
  - $h_{j+M+1}(X) = (X-\xi_j)_+^M,\ j = 1,\ldots,K$ where $x_+ = \max(0,x)$
- In practice the most widely used orders are $M = 0,1,3.$
- Parameter: **the order of the spline, # of knots and their placement**

Regression Splines

Natural splines

- We know that the behavior of polynomials fit to data tends to be erratic **near the boundaries**
- One way to remedy this problem is to force the piecewise polynomial to have a lower degree to the boundaries

Regression Splines

Natural splines

- A natural spline f of degree M with knots at $\xi_1 < \ldots < \xi_K$:
  - f is polynomial of degree M on each of $[\xi_1, \xi_2], \ldots, [\xi_{K-1}, \xi_K]$.
  - f is polynomial of degree $(M-1)/2$ on $(-\infty, \xi_1], [\xi_K, \infty)$.
  - $f^{(l)}$ is continuous at each of $\xi_1, \ldots, \xi_K$, for all $l = 0,...,M-1$
- # of parameters: $(M+1) \times (K-1) + ((M-1)/2 + 1) \times 2 - M \times K = K$

Regression Splines

Natural cubic splines

- Starting from the truncated power bases, we arrive at
- $N_1(X) = 1,\ N_2(X) = X, N_{k+2}(X) = d_k(X) - d_{K-1}(X)$, where
- $$d_k(X) = \frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k}$$

Regression Splines

Natural cubic splines

pf) $f(x) = \sum_{j=0}^{3} \beta_j X^j + \sum_{l=1}^{K} \beta_l (X - d_l)_+^3$ with constraint : linear at $(-\infty, d_1)$, (d_K, ∞)

$\Rightarrow$ for $x < d_1$, $f(x) = \sum_{j=0}^{3} \beta_j X^j$ i.e. $\beta_2 = \beta_3 = 0$. for $X > d_K$, $f(x) = \beta_0 + \beta_1 X + \sum_{l=1}^{K} \beta_l (X^3 - 3d_l X^2 + 3d_l^2 X - d_l^3)$

$\Rightarrow \sum_{l=1}^{K} \beta_l = \sum_{l=1}^{K} \beta_l d_l = 0$. $\Rightarrow$ Show that this is $\sum_l \theta_l N_l(X)$. $\Rightarrow \theta_0 = \beta_0$, $\theta_1 = \beta_1$

$f(x) = \beta_0 + \beta_1 X + \sum_{l=1}^{K} \beta_l (X - d_l)_+^3 = \theta_1 N_1(X) + \theta_2 N_2(X) + \sum_l \theta_{l+2} N_{l+2}(X)$.

$\Rightarrow \sum_l \theta_{l+2} \left[\frac{(X-d_l)_+^3 - (X-d_K)_+^3}{d_K - d_l} - \frac{(X-d_{K-1})_+^3 - (X-d_K)_+^3}{d_K - d_{K-1}} \right] = \sum_l \beta_l (X-d_l)_+^3 - (X-d_K)_+^3 \sum_l \beta_l - \frac{1}{d_K - d_{K-1}} \left[d_K \sum_l \beta_l - \sum_l \beta_l d_l \right] \left((X-d_{K-1})_+^3 - (X-d_K)_+^3 \right)$

$\theta_{l+2} = \beta_l (d_K - d_l)$

$= \sum_{l=1}^{K-2} \beta_l (X-d_l)_+^3 + (X-d_K)_+^3 (\beta_K + \beta_{K-1}) - \frac{1}{d_K - d_{K-1}} \left[d_K (\beta_K - \beta_{K-1}) + \beta_K d_K + \beta_{K-1} d_{K-1} \right] \left\{ (X-d_{K-1})_+^3 - (X-d_K)_+^3 \right\}$

$\beta_{K-1}(d_{K-1} - d_K)$

$= \sum_{l=1}^{K-2} \beta_l (X-d_l)_+^3 + (\beta_{K-1} + \beta_K)(X-d_K)_+^3 + \beta_{K-1}(X-d_{K-1})_+^3 - \beta_{K-1}(X-d_K)_+^3 = \sum_{l=1}^{K} \beta_l (X-d_l)_+^3$

Regression Splines

Natural cubic splines

- Example: South African Heart Disease
- The functional form of the model is:
  - logit $\left[Pr(\text{chd}\,|\,X)\right] = \theta_0 + h_1(X_1)^T\theta_1 + \ldots + h_p(X_p)^T\theta_p$
  - Note. This basis is not spline in $\mathbb{R}^p$

Smoothing splines

Smoothing splines estimator

- Consider the following problem:
  - $RSS(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt$ where λ is fixed parameter
    - $\lambda = 0$: f can be any function that interpolates the data
    - $\lambda = \infty$: the simple least squares line fit, since no second derivative can be tolerated

Smoothing splines

Smoothing splines estimator

- This criterion is defined on an **infinite-dimensional** function space which is a Sobolev space of functions for which the penalty term is defined.

- Remarkably, it can be shown that this problem has an explicit, **finite-dimensional, unique minimizer** which is **a natural cubic spline with knot at the unique values of the** $x_1, \ldots, x_N$ **:** $f(x) = N_1(x)\theta_1 + \ldots + N_N(x)\theta_N$ **.**

pf) For any interpolator f_0, we can always construct a natural cubic spline f s.t.
$\sum_{i=1}^{N}(y_i - f_0(x_i))^2 = \sum_{i=1}^{N}(y_i - f(x_i))^2$ i.e. we only consider second term.

Let $\tilde{g}$ be any differentiable on $[a,b]$ that interpolate N-pairs, g be natural cubic splines.

$\Rightarrow g$ is linear on $(-\infty, x_1) \cup (x_N, \infty)$ Let $h(x) = \tilde{g}(x) - g(x)$. then $h(x_1) = \cdots = h(x_N) = 0$.

$$\Rightarrow \int_a^b g''(x)h''(x)dx = [g''(x)h'(x)]_a^b - \int_a^b g^{(3)}(x)h'(x)dx = -\int_{x_1}^{x_N} g^{(3)}(x)h'(x)dx \quad (\because \int_a^{x_1} g^{(3)}(x)dx = \int_{x_N}^b g^{(3)}(x)dx = 0)$$

$\Rightarrow$ Since degree of g is 2, $g^{(3)}$ is constant on each (x_j, x_{j+1}). Let this d_j.

$$\Rightarrow \int_a^b g''(x)h''(x)dx = -\sum_{j=1}^{N-1}\int_{x_j}^{x_{j+1}} g^{(3)}(x)h'(x)dx = -\sum_{j=1}^{N-1} d_j[h(x_{j+1}) - h(x_j)] = 0.$$

$\Rightarrow$ which means $\int_a^b g''(x)^2dx = \int_a^b g''(x)\tilde{g}''(x)dx$.

$$\Rightarrow 0 \le \int_a^b h''(x)^2dx = \int_a^b g''(x)^2dx - 2\int_a^b g''(x)^2dx + \int_a^b \tilde{g}''(x)^2dx \Rightarrow \int_a^b g''(x)^2dx \le \int_a^b \tilde{g}''(x)^2dx$$

and equality holds if $h(x)$ is constant in $[a,b]$ but $h(x_j) = 0$ for $j = 1, \cdots, N$, so $h(x) = 0$ in $[a,b]$

Smoothing splines

Smoothing splines estimator

- Thus, plug the natural cubic spline estimator into the criterion, and find optimal θ.

- Define $\{N\}_{i,j} = N_j(x_i)$, $\left\{\Omega_N\right\}_{i,j} = \int N_i''(t)N_j''(t)dt$. Then we can rewrite as

  $$RSS(\theta, \lambda) = (\mathbf{y} - N\theta)^T(\mathbf{y} - N\theta) + \lambda\theta^T\Omega_N\theta$$

- $\hat{\theta} = (N^TN + \lambda\Omega_N)^{-1}N^T\mathbf{y}$; a generalized ridge regression.

Smoothing splines

Degree of freedom and Smoother matrix

- Suppose that λ is pre-chooses. This case is a example of a linear smoother because each $\hat{\theta}_j$ is linear combination of y_i.
- Let $\hat{\mathbf{f}} = N(N^T N + \lambda \Omega_N)^{-1} N^T \mathbf{y} = S_\lambda \mathbf{y}$ where S_λ is linear operator which is called smoother matrix.
- Now consider for M-cubic spline($M << N$), let B_ξ is expanded by bases at $x_1, \ldots, x_N$: $\hat{\mathbf{f}} = B_\xi (B_\xi^T B_\xi)^{-1} B_\xi^T \mathbf{y} = H_\xi \mathbf{y}$

Smoothing splines

Similarity & difference between S_λ and H_ξ

- Both are symmetric & positive semi-definite
  - Note. For $S_\lambda = N(N^T N + \lambda\Omega_N)^{-1}N^T$, enough to check that Ω_N is positive semidefinite. And

$$\forall\theta,\ \theta^T\Omega_N\theta = \sum_i\sum_j \theta_i \int N_i''(t)\int N_j''(t)dt\theta_j = \int(\sum_i N_i''(t)\theta_i)^2 dt \geq 0$$

- $H_\xi H_\xi = H_\xi$ (idempotent), while $S_\lambda S_\lambda + A = S_\lambda$ where A is positive semi-definite.

  This is a consequence of the **shrinking** nature of S_λ.

Smoothing splines

Similarity & difference between S_λ and H_ξ

- rank(H_ξ) = M, while rank(S_λ) = N.
- Note that $M = tr(H_\xi)$ gives the # of basis functions, which is # of parameters.
- By analogy we define **the effective degree of freedom** of a smoothing spline to be $\text{df}_\lambda = tr(S_\lambda)$.

Smoothing splines

Reinsch form

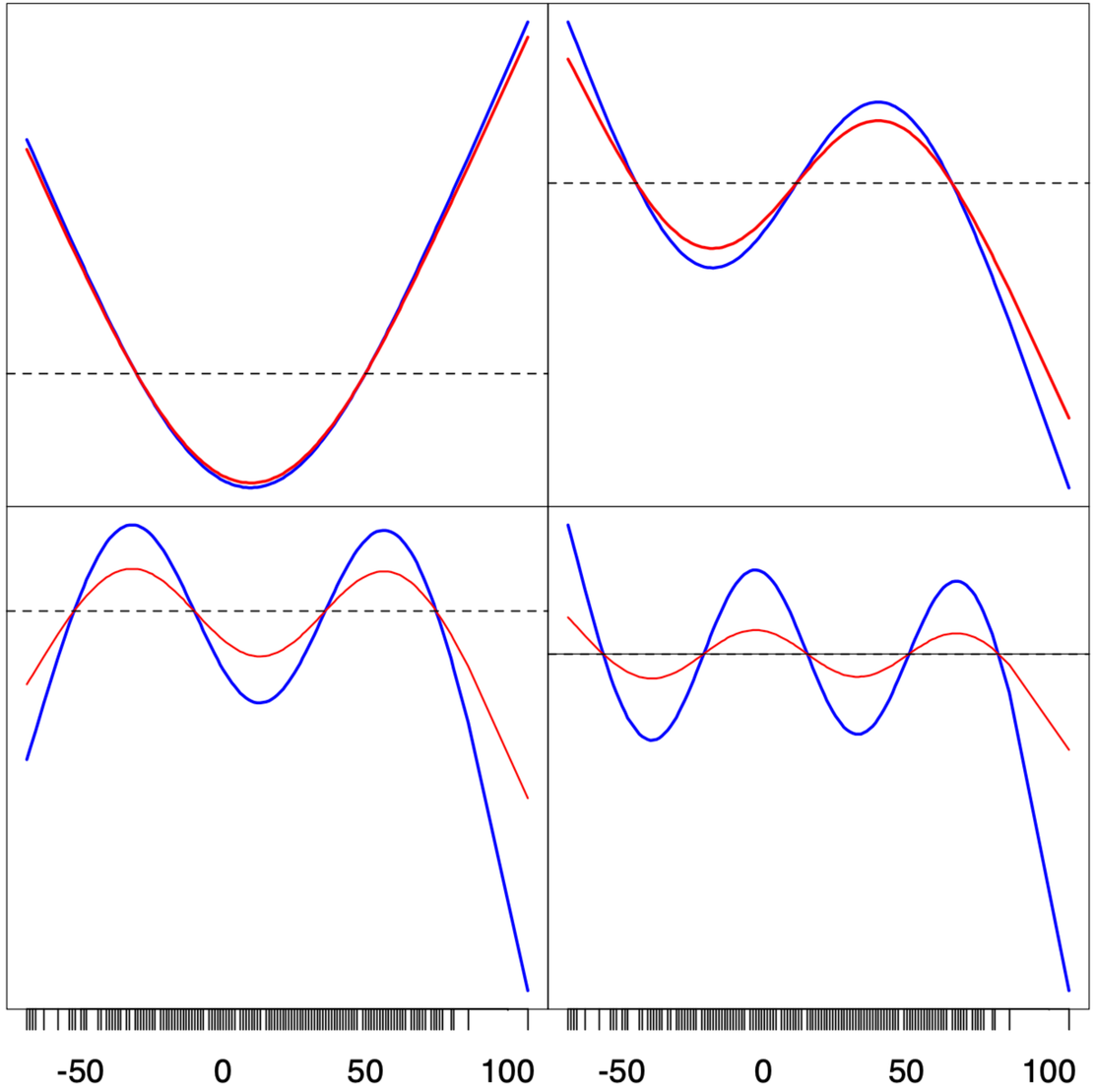
- The whole family of smoothing splines indexed by λ has same Eigen vectors.
- $\hat{\mathbf{f}} = \sum_{k=1}^{N} u_k(\rho_k(\lambda) u_k^T \mathbf{y})$ i.e. linear combination of $\{u_1, \ldots, u_N\}$ and

  differentially shrinking the contributions using $\rho_k(\lambda)$ which depends on λ,

  while $\hat{\mathbf{f}} = H_\xi \mathbf{y} = \sum_{k=1}^{N} h_i u_i u_i^T \mathbf{y}$ where h_i is either 0 or 1.

Smoothing splines

Reinsch form



Smoothing splines

Reinsch form

Pf. Since $N_1(x)=1$ & $N_2(x)=x$, $N_1''(x)=N_2''(x)=0$ i.e. $(\Omega_N)_{ij}=0$ if $i\le 2$ or $j\le 2$.

Note that $K=(N^T)^{-1}\Omega_N N^{-1}$ & for its eigenvalue d_k, S_λ's is $\frac{1}{1+\lambda d_k}$

$$\Rightarrow \Omega_N Y=0 \Rightarrow \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & & \end{bmatrix}\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \\ Y_3\Omega_{31}+\cdots+Y_N\Omega_{3N} \\ \vdots \\ Y_N\Omega_{N1}+\cdots+Y_N\Omega_{NN} \end{bmatrix}$$

$\Rightarrow$ i.e. $Y_1=\cdots=Y_N=0$. find v s.t. $N^{-1}v=(Y_1\ Y_2\ 0\ \cdots\ 0)^T$

Note that $N^TN=\begin{bmatrix} 1 & 0 & \\ 0 & 1 & \cdots \\ \vdots & \vdots & \\ 0 & 0 & \end{bmatrix}$ $\Rightarrow N^T\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}=\begin{bmatrix}1\\0\\ \vdots\\0\end{bmatrix}$ & $N^{-1}\begin{bmatrix}x_1\\ \vdots\\ x_N\end{bmatrix}=\begin{bmatrix}0\\1\\ \vdots\\0\end{bmatrix}$ $\Rightarrow v=C_1\begin{bmatrix}1\\ \vdots\\1\end{bmatrix}+C_2\begin{bmatrix}x_1\\ \vdots\\ x_N\end{bmatrix}=C_1 j+C_2 x$ for constant C_1, C_2.

$\Rightarrow$ Thus, nullity $(K)\ge 2 \Rightarrow \exists d_k$ s.t. $d_k=0$ & $\#k\ge 2 \Rightarrow 1$ be eigen value for S_λ & $\forall k\le 1$ ($\because \lambda>0,\ d_k\ge 0$)

$\Rightarrow$ the first two eigen values are always one & they correspond to the 2-dimensional eigen space of functions linear in x. & since $d_k=0$, linear functions are not penalized.

$\Rightarrow$ As $\lambda\to 0$, $df_\lambda=\sum_{k=1}^{N}\frac{1}{1+\lambda d_k}\to N$, and $S_\lambda\to VV^T=I$. & As $\lambda\to\infty$, $df_\lambda\to 2$ ($\because$ by 4), and $S_\lambda\to H$.

Smoothing splines

Equivalent kernel

- Recall that for test point x, smoother spline estimator is $\hat{f}(x) = (N_1(x), \ldots, N_N(x)^T(N^T N + \lambda \Omega_N)^{-1} N^T \mathbf{y} = w(x)^T \mathbf{y}$ where $w(x) = (w_1(x), \ldots, w_N(x))$ is a weight function.

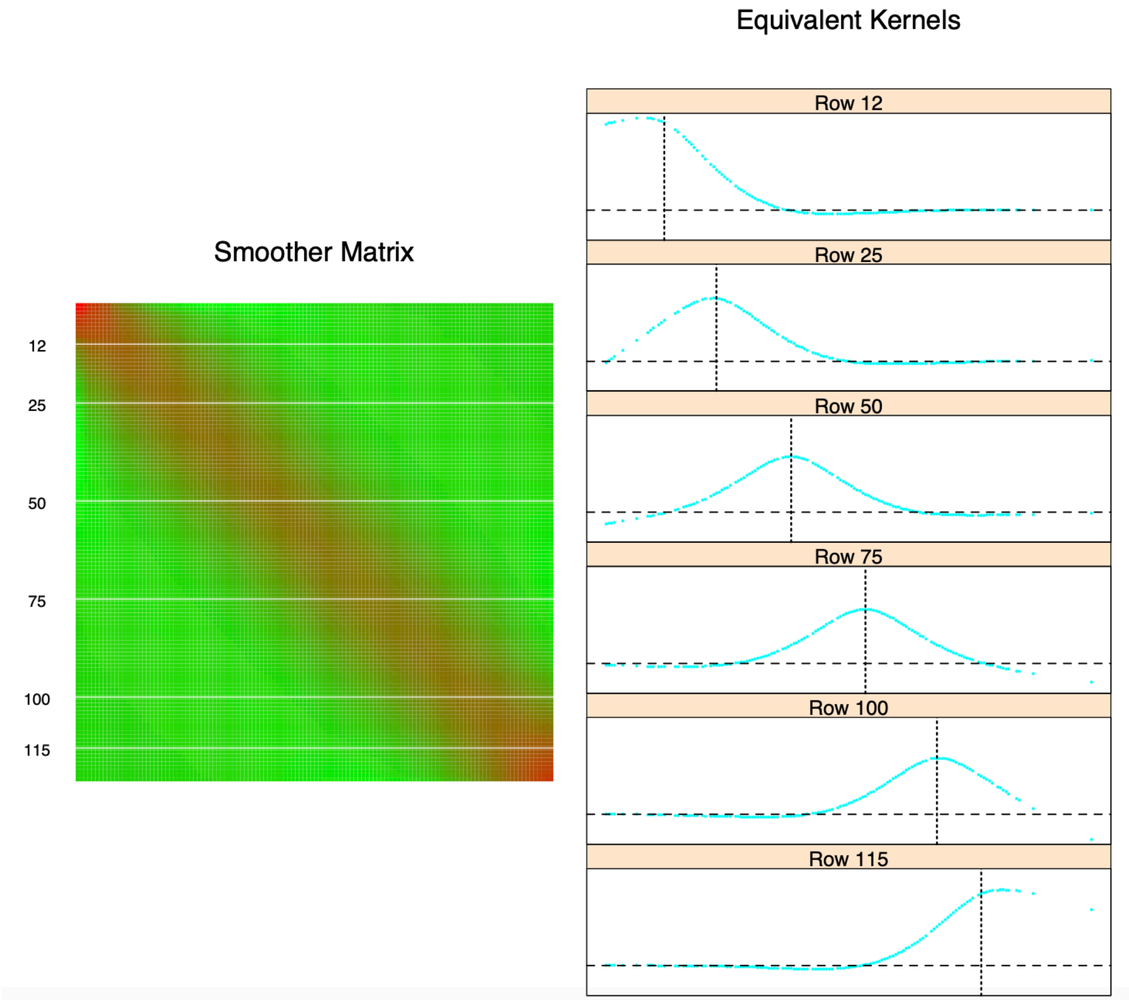
- Denote $w_i(x) = w(x, x_i)$ to emphasize that this weight has attribute to (x_i, y_i).

- Then, $\hat{f}(x) = w(x, x_1)y_1 + \ldots + w(x, x_n)y_N$.

Smoothing splines

Equivalent kernel

- Now consider for training data $x_1, \ldots, x_N$; i.e. $\hat{\mathbf{f}} = S_\lambda \mathbf{y}$
- i-th row of S_λ : $(w_1(x_i), \ldots, w_N(x_i))$
- The weights look like they are given by **translations of the same kernel**.



Smoothing splines

Equivalent kernel

- Indeed, under suitable regularity conditions, we have large N -approximation, $w(x,z) \approx \frac{1}{h(x)p(x)} K\left(\frac{x-z}{h(x}\right)$ where $p(x)$ is distribution at x, $h(x)$ is band-width $(\lambda/p(x))^{1/4}$ (adapts to local distribution of x).
- $K(t) = \frac{1}{2}\exp(-\frac{|t|}{\sqrt{2}})\sin(\frac{|t|}{\sqrt{2}} + \frac{\pi}{4})$ which is known as **Silberman kernel**.
- Thus, smoother spline estimator is asymptotically equivalent to kernel regression estimator

Multivariate splines

Splines

- There are no real multivariate extensions of spline estimator. In fact, even defining a multivariate spline is generally very tricky.
- Suppose a multivariate spline should be, like a univariate spline, a piecewise polynomial of degree k that is in C^{k-1}.
- k-th degree polynomial in d-dimension is the form:
  - $f(x) = \sum_{|\alpha| \leq k} \beta_\alpha \prod_{i=1}^{d} x_i^{\alpha_i}$ for coefficients β_α, multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$

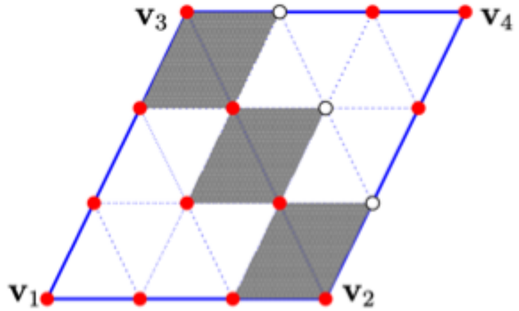
Multivariate splines

Splines

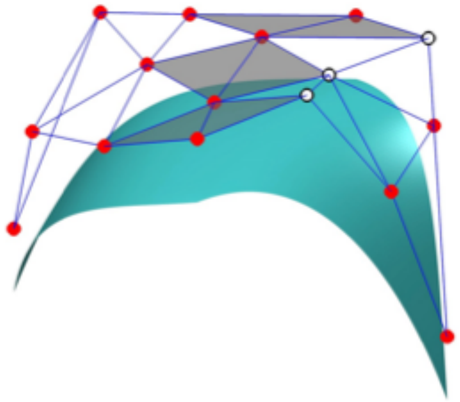
- Thus, (# of coefficients) = ${}_{d+1}H_k = {}_{d+k}C_k$
- Now consider for cubic degree in 2-dimension. It's tricky to construct C^2 piecewise cubic, when we consider the **"pieces" to be triangles**. Indeed, it's already hard to construct C^1 piecewise cubic.
- Suppose that we have two triangles sharing an edge e.

Multivariate splines

Splines



(a) Two domain triangles with C^1 constraints on Bézier ordinates.



(b) Two Bézier patches with C^1 continuity.

- Thus, (# of coefficients) = $_{2+1}H_3 =_5 C_2 = 10$ for each triangle, total 20-parameters.
- C^1 conditions:
  $\forall x \in e, \forall |\alpha| \leq 1, f^{(\alpha)}(x) = g^{(\alpha)}(x)$
  where f, g are two cubics on triangles
- The linear constraints are all **entangled.**

Multivariate splines

Tensor product Splines

- For any $f_1, \ldots, f_p$, define the tensor product $f_1 \otimes \ldots \otimes f_p$ as a function $\mathbb{R}^p \to \mathbb{R}$ by $f_1 \otimes \ldots \otimes f_p(x_1, \ldots, x_p) = \prod_{i=1}^{p} f_i(x_i)$

- Suppose $x \in \mathbb{R}^d$. Given a M-th degree spline basis with K-knots $\xi_1, \ldots, \xi_K$, which we denote by $h_1, \ldots, h_N$ $(N = M + K + 1)$. i.e. the function of the form: $f(x) = \sum_{j \in \{1,\ldots,N\}^d} \beta_j h_{j_1}(x_1) \otimes \ldots \otimes h_{j_d}(x_d)$

Multivariate splines

Tensor product Splines

- It is a polynomial on each hypercube of the form $[\xi_{j_1}, \xi_{j_1+1}] \times \ldots \times [\xi_{j_d}, \xi_{j_d+1}]$
- It is a tensor product of univariate M-th degree polynomials; i.e **not necessarily a M-th degree polynomial** in each hypercube.
- Since h_j be M-th degree polynomial, $f(x)$ has the form:

$$f(x) = \sum_{\alpha_1, \ldots \alpha_d \leq M} \beta_\alpha \prod_{i=1}^{d} x_i^{\alpha_i}$$

Multivariate splines

Thin plate Splines

- The idea is to set up the problem:

$$\min_{f \in \mathscr{H}} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_U \sum_{|\alpha|=m} \left[D^\alpha f(x)\right]^2 dx$$ for an integer $m \geq 1$, $U \subset \mathbb{R}^d$

  contains $x_1, \ldots, x_N$, $\mathscr{H}$ contains all functions f for the which the criterion is well-defined and finite. We can consider the Sobolev space $W^{m,2}(U)$.

- When $2m > d$, the problem is well-define, it admits a solution. This solution is called the thin plate spline estimator. (Sobolev embedding theorem).

Multivariate splines

Sobolev Space

- Let $f \in L_1([a,b])$ (which means $\int_a^b |f(x)|dx < \infty$). We say that

  $g \in L_1([a,b])$ is **weak derivatives of** f if $\int_a^b f(x)\rho'(x)dx = -\int_a^b g(x)\rho(x)dx$

  for all $\rho \in C^\infty$ with $\rho(a) = \rho(b) = 0$

Multivariate splines

Sobolev Space

- Motivation: For differentiable f, ρ, $\int_a^b f(x)\rho'(x) + f'(x)\rho(x)dx = \left[f(x)\rho(x)\right]_b^a = 0$ i.e. $g = f'$
- g is unique almost everywhere, and denoted by Df, called weak derivative of f.
- Ex. $f(x) = x_+$ with $Df(x) = I(x > 0)$.
  - $\int_{-\infty}^{\infty} x_+\rho'(x)dx = \int_o^{\infty} x\rho(x)dx = \left[x\rho(x)\right]_0^{\infty} - \int_0^{\infty} \rho(x)dx = -\int_{-\infty}^{\infty} I(x > 0)\rho(x)dx$

Multivariate splines

Sobolev Space

- For higher-order derivatives, denote $D^{\alpha}\phi = \frac{\partial^{|\alpha|}\phi}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ where

  $|\alpha| = \alpha_1 + \ldots + \alpha_d$. Then f is said to have α-th weak differentiable if $\int_U f(x) D^{\alpha}\rho(x)dx = -(1)^{|\alpha|} \int_U g(x)\phi(x)dx$ for all $\phi \in C_c^{\infty}(U)$ where $U \subset \mathbb{R}^d$ is an open set, $C_c^{\infty}(U)$ is the set of all infinitely differentiable functions with compact support on U .

- Again, g is unique almost everywhere, and we denote by $D^{\alpha}f$.

Multivariate splines

Sobolev Space

- For $k \in \mathbb{Z}_+,\ p \in (0,\infty),$ and open domain $U \subset \mathbb{R}^d$, we define the Sobolev space: $W^{k,p}(U) = \left\{ f : U \to \mathbb{R} : \forall\, |\alpha| \leq k,\ \|D^\alpha f\|_p < \infty \right\}$ where

$$\|f\|_p = \left(\int_U |f(x)|^p \, dx \right)^{1/p}$$

Multivariate splines

Return to Thin plate splines

- For $2m > d$, which we called **the super critical regime**, $W^{m,2}$ embeds continuously into Hoder space $C^{p+\gamma}(U)$ (which contains all functions f s.t. the α-th derivative of f is bounded for all $|\alpha| \leq p + \gamma$ with L^∞ norm and Lipschitz for $|\alpha| = p + \gamma$). Thus, $W^{m,2}$ is also continually into $C^0(U)$ (which is continuous functions on U with L^∞ norm).

- Therefore, for sequence of function f_k s.t. $\|f_k - f\|_{W^{m,2}(U)}$ as $k \to \infty$, this implies $\|f_k - f\|_{L^\infty}$, which implies uniform convergence i.e. $\forall x \ f_k(x) \to f(x)$

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

Multivariate splines

Thin plate Splines

- In $\mathbb{R}^2$, it is valid to take $m = 2$.
- $$\min_{f \in \mathcal{H}} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f(x}{\partial x_1^2} \right) + \left(\frac{\partial^2 f(x}{\partial x_1 x_2} \right) + \left(\frac{\partial^2 f(x}{\partial x_2^2} \right) \right]^2 dx$$
- The solution has the form: $f(x) = \beta_0 + \beta^T x + \sum_{j=1}^{N} \alpha_j h_j(x)$ where

  $h_j(x) = \|x - x_j\|^2 \log \|x - x_j\|$ which is example of radial basis function.

Multivariate splines

Thin plate Splines

- For a penalty term to be finite, a necessary and sufficient condition on the coefficients is: $\sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \alpha_i x_i = 0$ i.e. 3 linear constraints ($x_i \in \mathbb{R}^2$)

- Therefore, we have $(1 + 2 + n) - 3 = n$ free parameters.

Reproducing Kernel Hilbert spaces

Hilbert spaces

- Let $\mathscr{H}$ be vector space over $\mathbb{R}$. An inner product on $\mathscr{H}$ is a function $\mathscr{H} \times \mathscr{H} \to \mathbb{R}$, denoted $\langle x, y \rangle_{\mathscr{H}}$, s.t. $\forall f, g, h \in \mathscr{H}$, $\forall c \in \mathbb{R}$, the following hold:
  - $\langle f + h, g \rangle_{\mathscr{H}} = \langle f, g \rangle_{\mathscr{H}} + \langle h, g \rangle_{\mathscr{H}}$
  - $\langle cf, g \rangle_{\mathscr{H}} = c \langle f, g \rangle_{\mathscr{H}}$
  - $\langle f, g \rangle_{\mathscr{H}} = \langle g, f \rangle_{\mathscr{H}}$
  - $\langle f, f \rangle_{\mathscr{H}} > 0$ if $f \neq 0$

Reproducing Kernel Hilbert spaces

Hilbert spaces

- We can always define a norm based on inner product, denoted $\|\cdot\|_{\mathscr{H}}$ by $\|f\|_{\mathscr{H}}^2 = \langle f, f \rangle_{\mathscr{H}}$
- An inner product space $\mathscr{H}$ is called a Hilbert space if it is complete. ($\|f_m - f_n\|_{\mathscr{H}} \to 0$ as $m, n \to \infty$)
- Note.
  - Any Hilbert space is Banach space which is complete norm space
  - $f_n \to f$ does not imply that $f_n(x) \to f(x)$ for all x (It's true for RKHS)

Reproducing Kernel Hilbert spaces

Kernels

- A kernel is function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a map $\phi : \mathcal{X} \to \mathcal{H}$ s.t. $\forall x, y \in \mathcal{X},\ k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$
- By definition of inner products, k is symmetric & non-negative.
- For $x_1, \ldots, x_N \in \mathcal{X}$, define $K \in \mathbb{R}^{N \times N}$ with $\{K\}_{ij} = k(x_i, x_j)$. Then for any $a \in \mathbb{R}^N$, $a^T K a = \sum_{i,j} a_i K(x_i, x_j) a_j = \sum_{i,j} \left\langle a_i \phi(x_i), a_j \phi(x_j) \right\rangle \geq 0$ i.e. K is positive semidefinite.
- A function k is called positive semi-definite.

Reproducing Kernel Hilbert spaces

Kernels

- Remarkably, any positive semidefinite $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel, which means **we don't have to find ϕ.**
- Ex. Let $\mathcal{X} = \mathbb{R}^d$
  - Polynomial kernel: $k(x, y) = (1 + x^T y)^m$
  - Exponential kernel: $k(x, y) = \exp(x^T y)$
  - Gaussian kernel: $k(x, y) = \exp(-\|x - y\|_2^2 / \sigma^2)$

Reproducing Kernel Hilbert spaces

RKHS

- A Hilbert space of function on $\mathcal{X}$, denoted $\mathcal{H}$ is called a representing kernel Hilbert space with kernel k, the following hold:
  - $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$ **(reproducers of evaluation)**
  - $\forall f \in \mathcal{H}, \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ for all $x \in \mathcal{X}$ **(reproducing property)**
- By reproduction property,
  $\forall x, y \in \mathcal{X}, \ \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = k(y, x) = k(x, y)$ i.e. $\phi(x) = k(\cdot, x)$

Reproducing Kernel Hilbert spaces

RKHS

- For $x \in \mathcal{X}$, denoted by $\delta_x : \mathcal{H} \to \mathbb{R}$ the corresponding **evaluation operator** on $\mathcal{H}$ (i.e. $\delta_x(f) = f(x)$).

- Then a Hilbert space of functions $\mathcal{H}$ is an RKHS iff all of the evaluation operators are continuous ($\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\exists M < \infty$ s.t. $|\delta_x(f)| = |f(x)| \leq M\|f\|_{\mathcal{H}}$)

Reproducing Kernel Hilbert spaces

RKHS

- **Dual Space**: Let $\mathscr{H}$ be Hilbert space over $\mathbb{R}$. We define the dual space of $\mathscr{H}$ to be $\mathscr{L}(\mathscr{H}, \mathbb{R}) = \left\{L : \mathscr{H} \to \mathbb{R} : L \text{ is continuous linear operator}\right\}$, denoted $\mathscr{H}^*$

- **Riesz Representation Theorem:** Let $\mathscr{H}$ be Hilbert space over $\mathbb{R}$. $\forall \rho \in \mathscr{H}^*, \exists ! f_\rho \in \mathscr{H}$ called the Riesz representation of ρ s.t. $\rho(x) = \left\langle x, f_\rho \right\rangle_{\mathscr{H}}$ for all $x \in \mathscr{H}$

pf: (existence): Let $K = N(\varphi) = \{m \in \mathcal{H} : \varphi(m) = 0\}$. If $K = \mathcal{H}$, then let $f_\varphi = 0$.

Now consider for $K \subset \mathcal{H}$. The continuity of φ implies that K is closed subspace of $\mathcal{H}$

($\because K = \varphi^{-1}(\{0\})$ & $\{0\}$ is a closed subset of $\mathbb{F}$).

$\Rightarrow \mathcal{H}$ can be written as $\mathcal{H} = K \oplus K^\perp$. $\Rightarrow$ Since $K \neq \mathcal{H}$, $K^\perp \neq \emptyset$. i.e. exists non-zero $p \in K^\perp$.

$\Rightarrow \forall h \in \mathcal{H}$. $\varphi[(\varphi h)p - (\varphi p)h] = (\varphi h)(\varphi p) - (\varphi p)(\varphi h) = 0$ i.e. $(\varphi h)p - (\varphi p)h \in K$.

$\Rightarrow$ Since $p \in K^\perp$, $\langle p, (\varphi h)p - (\varphi p)h \rangle_{\mathcal{H}} = (\varphi h)\|p\|_{\mathcal{H}}^2 - (\varphi p)\langle p, h \rangle_{\mathcal{H}}$

$\Rightarrow \varphi h = \dfrac{(\varphi p)\langle p, h \rangle_{\mathcal{H}}}{\|p\|_{\mathcal{H}}^2}$ for $\forall h \in \mathcal{H}$. Therefore, let $f_\varphi = \dfrac{\langle p, h \rangle_{\mathcal{H}}}{\langle p, p \rangle_{\mathcal{H}}}(\varphi p)$

(uniqueness): Suppose $f, g \in \mathcal{H}$. are s.t. $\varphi(z) = \langle z, f \rangle_{\mathcal{H}}$ and $\varphi(z) = \langle z, g \rangle_{\mathcal{H}}$.

$\Rightarrow \langle z, f - g \rangle_{\mathcal{H}} = \langle z, f \rangle_{\mathcal{H}} - \langle z, g \rangle_{\mathcal{H}} = \varphi(z) - \varphi(z) = 0$ for $\forall z \in \mathcal{H}$.

$\Rightarrow \langle \cdot, f - g \rangle_{\mathcal{H}}$ is constant 0, which implies $0 = \|g - f\|_{\mathcal{H}} \Rightarrow f = g$.

Reproducing Kernel Hilbert spaces

RKHS

- The evaluation operator is linear:
  - $\forall f, g \in \mathscr{H},\ \delta_x(f+g) = (f+g)(x) = f(x) + g(x) = \delta_x(f) + \delta_x(g)$
  - $\forall a \in \mathbb{R},\ \forall f \in \mathscr{H},\ \delta_x(af) = af(x) = a\delta_x(f)$
- Therefore, if all of the evaluation operators are continuous, then it is element of $\mathscr{H}^*$. Then by Reisz representation theorem, the reproducing property holds.

Reproducing Kernel Hilbert spaces

RKHS

- Recall that Sobolev space $W^{m,2}(U)$ (where $U \subset \mathbb{R}^d$) is Hilbert space and it's continuous iff $2m > d$.
- Thus, $W^{m,2}(U)$ is an RKHS iff $2m > d$. Then, what is its kernel?
- For $U = \mathbb{R}^d$, $k(x, y) = \int \frac{\exp(2\pi i(x-y)^T u}{1 + \sum_{0 \leq |\alpha| \leq m} \prod_{j=1}^{d} (2\pi u_j)^{2\alpha_j}}$

Reproducing Kernel Hilbert spaces

RKHS

- For $d = 1, m = 2$, we get the smoothing spline kernel:

$$k(x, y) = \frac{1}{\sqrt{3}} \exp\left(-\frac{\sqrt{4}\,|x - y|}{2}\right) \sin\left(\frac{|x - y|}{2} + \frac{\pi}{6}\right)$$

- For $d = 2, m = 2$, we get the tt: $k(x, y) = \frac{1}{16\pi} \|x - y\|_2^2 \log \|x - y\|_2^2$

Reproducing Kernel Hilbert spaces

RKHS

- **Representer Theorem**: Let $\mathscr{H}$ be RKHS, and consider following problem $\underset{f \in \mathscr{H}}{\text{argmin}} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|_{\mathscr{H}}^2$ which is **infinite dimensional** problem. Then the unique solution $f(x) = \sum_{i=1}^{N} c_i k(x, x_i)$ where $c_1, \ldots, c_N \in \mathbb{R}$ which is **finite dimensional**.

pf) Let $\mathcal{H}_0 = \text{span}\{K(\cdot, x_1), \cdots, K(\cdot, x_n)\}$. $\mathcal{H}_1 = \{f \in \mathcal{H} : f(x_i) = 0, i = 1, \cdots, n\}$.

$\Rightarrow$ By reproducing property, $\forall f \in \mathcal{H}_1$, $\langle f, K(\cdot, x_i)\rangle_{\mathcal{H}} = f(x_i) = 0$ for $\forall i$.

$\Rightarrow \forall g \in \mathcal{H}_0$. g can be written as $g = \sum_{j=1}^{n} c_j K(\cdot, x_j)$

$\Rightarrow \forall f \in \mathcal{H}_1, \forall g \in \mathcal{H}_0$, $\langle f, g\rangle_{\mathcal{H}} = \langle f, \sum_{j=1}^{n} c_j K(\cdot, x_j)\rangle = 0$ for $\forall c_j \in \mathbb{R}$.

$\Rightarrow \mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ i.e. $\forall f \in \mathcal{H}$. $f = f_0 + f_1$ where $f_0 \in \mathcal{H}_0$, $f_1 \in \mathcal{H}_1$

$\Rightarrow \|f\|_{\mathcal{H}}^2 = \|f_0 + f_1\|_{\mathcal{H}}^2 = \|f_0\|_{\mathcal{H}}^2 + \|f_1\|_{\mathcal{H}}^2$ & $f_1(x_i) = 0$ for $\forall i$ ($\because f_1 \in \mathcal{H}_1$)

$\Rightarrow f(x_i) = f_0(x_i)$ for $\forall i$. $\Rightarrow \sum_{i=1}^{N} L(y_i, f(x_i)) = \sum_{i=1}^{N} L(y_i, f_0(x_i))$

$\Rightarrow RSS_f = \sum_{i=1}^{N} L(y_i, f_0(x_i)) + \lambda\|f_0\|_{\mathcal{H}}^2 + \lambda\|f_1\|_{\mathcal{H}}^2 \geq RSS_{f_0}$ ($\lambda\|f_1\|_{\mathcal{H}}^2 \geq 0$)

$\therefore \hat{f} \in \text{span}\{K(\cdot, x_1), \cdots, K(\cdot, x_n)\} \Rightarrow \hat{f}(x) = \sum_i c_i K(x, x_i)$.

Reproducing Kernel Hilbert spaces

RKHS

- Now, $f(x) = \sum_{i=1}^{N} c_i k(x, x_i)$ into criterion,

  $\|f\|^2_{\mathscr{H}} = \sum_{i,j} c_i c_j \left\langle k(\cdot, x_i), k(\cdot, x_j) \right\rangle_{\mathscr{H}} = \sum_{i,j} c_i c_j k(x_i, x_j)$ by reproducing property.

- Thus we reduce to finite dimensional: $\min_c L(\mathbf{y}, Kc) + \lambda c^T K c$

- The machinery above is driven by the choice of kernel k and the loss function L