Reproducing Kernel Hilbert Space

Ergan Shang, Yan Chen

USTC

November 27, 2022

Overview

Overview

Theorem 12.5 (Riesz representation theorem)

let L be a bounded linear functional on a Hilbert space. Then there exists a unique $g \in \mathcal{H}$ such that $L(f) = \langle f, g \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. And we refer to g as the representer of the functional L.

Also recall the definitions of inner product and Hilbert space.

Overview

PSD kernel function

Definition 12.6

A symmetric bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ matrix with elements $K_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semidefinite.

12.7 Let $\mathcal{X} = \mathbb{R}^d$, we define $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. Then with $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, we have

$$\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} = \sum_{i,j=1}^n \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \left\| \sum_{i=1}^n \alpha_i \boldsymbol{x}_i \right\|_2^2 \geq 0.$$

Polynomials kernels and Feature maps

12.8 We let $\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle)^2$, so that

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^{d} x_j^2 z_j^2 + 2 \sum_{i<j} x_i x_j z_i z_j.$$

Setting $D = d + \binom{d}{2}$, and define **feature mapping** $\phi : \mathbb{R}^d \to \mathbb{R}^D$ with entries

$$\phi(\boldsymbol{x}) = \begin{bmatrix} x_j^2, & \text{for } j = 1, 2, \cdots, d \\ \sqrt{2} x_i x_j, & \text{for } i < j \end{bmatrix}.$$

As a result, we can write $\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z}) \rangle_{\mathbb{R}^D}$, which is PSD followed by the last example.

Feature maps

12.10 Consider the Fourier basis $\phi_j(x) = \sin\left(\frac{(2j-1)\pi x}{2}\right)$ and $\langle \phi_j, \phi_k \rangle = \int_0^1 \phi_j(x)\phi_k(x)dx = \delta_{jk}$. Given some sequence $\{\mu_j\}_{j=1}^{\infty}$ with $\sum_{j=1}^{\infty} \mu_j < \infty$, we let the feature map as

$$\Phi(x) := (\sqrt{\mu_1}\phi_1(x), \sqrt{\mu_2}\phi_2(x), \cdots).$$

By constrcution

$$\|\Phi(x)\|^2 = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) \leq \sum_{j=1}^{\infty} \mu_j < \infty \Rightarrow \Phi(x) \in \ell^2(\mathbb{N}).$$

Therefore, we define $\mathcal{K}(x, z) = \langle \Phi(x), \Phi(z) \rangle_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j(x)\phi_j(z)$ is a PSD kernel.

Gaussian kernel

12.9 Choose $\mathcal{X} \subset \mathbb{R}^d$ and consider Gaussian kernel $\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$. In order to prove it is a PSD, we can first expand exp into polynomials and by first proving product of PSD is also a PSD, and then limit of summation of PSD is also a PSD. To be specific,

(1) Let $\mathcal{K}(x, z)$ is PSD, then the polynomials $P(\mathcal{K}(x, z))$ is also a PSD, where $P(x) = \sum_{i=0}^{n} a_i x^i$.

(2) We consider $\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right) \cdot \exp\left(\langle \boldsymbol{x}, \boldsymbol{z}\rangle/\sigma^2\right) \cdot \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$.

We only need to prove $\mathcal{K}(x, y) = \mathcal{K}_1(x, y)\mathcal{K}_2(x, y)$ is a PSD whenever $\mathcal{K}_1$ and $\mathcal{K}_2$ are PSDs. By Linear Algebra, we let $C = A^\top A$ and $D = B^\top B$ where $c_{ij} = \mathcal{K}_1(x_i, x_j)$ and $d_{ij} = \mathcal{K}_2(x_i, x_j)$. By defining $A = (a_1, \cdots, a_n)$ and $B = (b_1, \cdots, b_n)$, we have $c_{ij} = \sum_k a_{ik}a_{jk}$ and $d_{ij} = \sum_\ell b_{i\ell}b_{j\ell}$. By denoting $e_{ij} = \mathcal{K}_1(x_i, x_j)\mathcal{K}_2(x_i, x_j)$, we calculate

$$\boldsymbol{u}^\top E\boldsymbol{u} = \sum_{i,j} u_i u_j e_{ij} = \sum_{i,j}\sum_{k,\ell} u_i u_j a_{ik}a_{jk}b_{i\ell}b_{j\ell}$$

$$= \sum_{k,\ell}\left(\sum_i u_i a_{ik}b_{i\ell}\right)\left(\sum_j u_j a_{jk}b_{j\ell}\right) = \sum_{k,\ell}\left(\sum_i u_i a_{ik}b_{i\ell}\right)^2 \geq 0.$$

Outline

Constructing from a PSD

We propose the RKHS constructed from a PSD has the following property

$$\langle f, \mathcal{K}(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}} = f(\boldsymbol{x}) \quad \forall f \in \mathcal{H},$$

which is known as the kernel reproducing property. Given functions of form $f(\cdot) = \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j)$ and $\bar{f} = \sum_{k=1}^{\bar{n}} \bar{\alpha}_k \mathcal{K}(\cdot, \bar{x}_k)$, by the linearity of inner product, we have

$$\langle f, \bar{f}\rangle = \sum_{j=1}^{n}\sum_{k=1}^{\bar{n}} \alpha_j \bar{\alpha}_k \mathcal{K}(x_j, \bar{x}_k),$$

and moreover, this kind of inner product satisfies the reproducing property by

$$\langle f, \mathcal{K}(\cdot, x)\rangle = \sum_{j=1}^{n} \alpha_j \mathcal{K}(x_j, x) = f(x).$$

Explicit Version

Theorem 12.11

Given any PSD $\mathcal{K}$, there is a unique Hilbert space $\mathcal{H}$ in which the kernel satisfies the reproducing property. It is known as the reproducing kernel Hilbert space associated with $\mathcal{K}$.

Proof:

About the sensibility of inner product, we only need to prove $\|f\|_{\mathcal{H}}^2 = 0$ iff $f = 0$. We suppose that $\langle f, f\rangle_{\tilde{\mathcal{H}}} = \sum_{i,j=1}^{n} \alpha_i \alpha_j \mathcal{K}(x_i, x_j) = 0$, then by arbitrarily choosing $a \in \mathbb{R}$, we have

$$0 \leq \|a\mathcal{K}(\cdot, x) + \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x_i)\|^2 = a^2 \mathcal{K}(x, x) + 2a \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i).$$

Since $\mathcal{K}(x, x) \geq 0$ and $a \in \mathbb{R}$ is arbitrary, we have $f(x) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(x, x_i) = 0$.

Proof

Then we need to make a complete inner product space, i.e. the Hilbert space. Assuming $\{f_n\}_{n=1}^{\infty}$ is a Cauchy sequence, then $\{f_n(x)\}_{i=1}^{\infty} \subset \mathbb{R}$ is a Cauchy sequence, so that we define $f(x) = \lim_{n\to\infty} f_n(x)$ and also $\|f\|_{\mathcal{H}} := \lim_{n\to\infty} \|f_n\|_{\tilde{\mathcal{H}}}$. To verify it is well-defined, we have to prove that when the Cauchy sequence $\{g_n\}_{n=1}^{\infty}$ in $\tilde{\mathcal{H}}$ such that $\lim_{n\to\infty} g_n(x) = 0 \quad \forall x \in \mathcal{X}$, we also have $\lim_{n\to\infty} \|g_n\| = 0$. Otherwise, there is a subsequence such that $\lim_{n\to\infty} \|g_n\|^2 = 2\epsilon > 0$, so that for m, n large enoug, we have $\|g_n\|^2 \geq \epsilon$ and $\|g_m\|^2 \geq \epsilon$ and also $\|g_n - g_m\| \leq \epsilon/2$. Then we write $g_m(\cdot) = \sum_{i=1}^{N_m} \alpha_i \mathcal{K}(\cdot, x_i)$. By reproducing property,

$$\langle g_m, g_n \rangle = \sum_{i=1}^{N_m} \alpha_i \langle \mathcal{K}(\cdot, x_i), g_n \rangle = \sum_{i=1}^{N_m} \alpha_i g_n(x_i) \to 0.$$

By

$$\|g_n - g_m\|^2 = \|g_n\|^2 + \|g_m\|^2 - 2\langle g_n, g_m \rangle,$$

we get contradiction.

Proof

Finally, we prove the uniqueness. Suppose that $\mathbb{G}$ is another Hilbert space with $\mathcal{K}$ being its kernel. Since $\mathbb{G}$ is complete and closed under linear operations, we have $\mathcal{H} \subset \mathbb{G}$, so that $\mathbb{G} = \mathcal{H} \bigoplus \mathcal{H}^{\perp}$. Let $g \in \mathcal{H}^{\perp}$ and noting $\mathcal{K}(\cdot, x) \in \mathcal{H}$, then $g(x) = \langle \mathcal{K}(\cdot, x), g \rangle_{\mathbb{G}} = 0$. We conclude that $\mathcal{H}^{\perp} = \{0\}$, thus $\mathcal{H} = \mathbb{G}$.

Outline

Abstract Version

Observing that $f(x) = \langle f, \mathcal{K}(\cdot, x)\rangle$, we can view $\mathcal{K}(\cdot, x)$ as the evaluation function $L_x : \mathcal{H} \to \mathbb{R}$ that performs $f \mapsto f(x)$. By Riesz representation theorem, it means all evaluation functions in RKHS are bounded. A direct question is that how large the class of Hilbert space where the evaluation functions are bounded is?

Definition 12.12

A reproducing kernel Hilbert space $\mathcal{H}$ is a Hilbert space of real-valued functions on $\mathcal{X}$ such that for each $x \in \mathcal{X}$, the evaluation functional $L_x : \mathcal{H} \to \mathbb{R}$ is bounded, i.e., there exists some $M < \infty$ such that $|L_x(f)| \leq M\|f\|$ for all $f \in \mathcal{H}$.

Theorem 12.13

Given and Hilbert spaec $\mathcal{H}$ in which the evaluation functionals are all bounded, there is a unique PSD kernel $\mathcal{K}$ that satisfies the reproducing property.

Proof:

By Riesz representation theorem, there exists some element $R_x \in \mathcal{H}$ such that $f(x) = L_x(f) = \langle f, R_x \rangle$ for all $f \in \mathcal{H}$. We define $\mathcal{K}$ via $\mathcal{K}(x, z) = \langle R_x, R_z \rangle$. We only need to show it is positive semidefinite. In fact

$$\boldsymbol{\alpha}^\top K \boldsymbol{\alpha} = \sum_{j,k=1}^{n} \alpha_j \alpha_k \mathcal{K}(x_j, x_k) = \left\langle \sum_{j=1}^{n} \alpha_j R_{x_j}, \sum_{j=1}^{n} \alpha_j R_{x_j} \right\rangle = \left\| \sum_{j=1}^{n} \alpha_j R_{x_j} \right\|_{\mathcal{H}}^2 \geq 0.$$

Proof

It remains to prove the reproducing property: by

$$\mathcal{K}(y, x) = \langle R_y, R_x \rangle = R_x(y)$$

, we can see that $\mathcal{K}(\cdot, x) = R_x(\cdot)$, then by definition

$$f(x) = \langle f, R_x \rangle = \langle f, \mathcal{K}(\cdot, x) \rangle,$$

which is the reproducing property.
Finally, if there exists another kernel $\tilde{\mathcal{K}}$ satisfying the properties above, we can see

$$\begin{aligned}\mathcal{K}(x, x') &= \langle \mathcal{K}(\cdot, x), \mathcal{K}(\cdot, x') \rangle = \langle \mathcal{K}(\cdot, x), \tilde{\mathcal{K}}(\cdot, x') \rangle \\ &= \langle \tilde{\mathcal{K}}(\cdot, x), \tilde{\mathcal{K}}(\cdot, x') \rangle = \tilde{\mathcal{K}}(x, x') \quad \forall x, x' \in \mathcal{X}.\end{aligned}$$

Examples

12.14 We let $\mathcal{K}(x,z)=\langle x,z\rangle$ and $f(x)=\sum_j \alpha_j \mathcal{K}(x,x_j)$, thus

$$f(x)=\langle f,\mathcal{K}(\cdot,x)\rangle=\left\langle \sum_j \alpha_j \mathcal{K}(\cdot,x_j),\mathcal{K}(\cdot,x)\right\rangle=\sum_j \alpha_j \mathcal{K}(x,x_j)$$

$$=\sum_j \alpha_j\langle x,x_j\rangle=\left\langle x,\sum_j \alpha_j x_j\right\rangle.$$

It means the evaluation functional has the form $z\mapsto\langle z,\sum_{i=1}^n \alpha_i x_i\rangle$, i.e., $f_\beta(\cdot)=\langle\cdot,\beta\rangle$ and the inner product in RKHS is formed by $\langle f_\beta,f_{\beta'}\rangle_{\mathcal{H}}=\langle\beta,\beta'\rangle$.

Examples

12.16 (A simple Sobolev space) Consider the functions

$$\mathbb{H}^1[0,1] := \{f : [0,1] \to \mathbb{R} | f(0) = 0, \text{and } f \text{ is absolutely continuous with } f' \in L^2[0,1]\}.$$

The inner product is defined as $\langle f, g\rangle := \int_0^1 f'(z)g'(z)dz$. In order to prove it is an RKHS, we claim its representer of evaluation: $R_x(z) = \min\{x, z\} \Rightarrow R_x'(z) = \mathbb{I}_{[0,x]}(z)$. We can calculate

$$\langle f, R_x\rangle = \int_0^1 f'(z)R_x'(z)dz = \int_0^x f'(z)dz = f(x)$$

by absolutely continuous.
Also by the process of the proof of theorem 12.13, we know that the PSD $\mathcal{K}(x, z) = \langle R_x, R_z\rangle = \int_0^1 \mathbb{I}_{[0,x]}(t)\mathbb{I}_{[0,z]}(t)dt = \langle \mathbb{I}_{[0,x]}, \mathbb{I}_{[0,z]}\rangle_{L^2[0,1]}$, therefore providing a Gram representation, leading to positive semidefinite. We conclude that $\mathcal{K}(x, z) = \min\{x, z\}$ is the unique PSD kernel.

RMK: RKHS ensures that convergence of a sequence of functions in RKHS implies pointwise convergence, which means: if we let $f_n \to f^*$ in $\mathcal{H}$ norm, we have

$$|f_n(x) - f^*(x)| = |\langle f_n - f^*, R(\cdot, x)\rangle| = |L_x(f_n - f^*)| \leq \|L_x\|\|f_n - f^*\| \to 0.$$

12.15 ($L^2[0,1]$ is not an RKHS) Consider the sequence of functions $f_n(x) = x^n$ and since $\int_0^2 f_n^2(x)dx = \frac{1}{2n+1} \to 0$. Therefore, $\|f_n\|_{\mathcal{H}} \to 0$. However, $f_n(1) \equiv 1$, thus not pointwise convergent.

Overview

Motivations

12.18 (PSD matrices) Let $\mathcal{X} = [d]$ be equipped with Hamming metric, i.e. $P(\{j\}) = 1$ be the counting measure on this discrete space. Define a PSD kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the matrix $K = (K_{ij}) = (\mathcal{K}(i,j))_{i,j=1}^n$. Define the integral operator as

$$T_{\mathcal{K}}(f)(x) = \int_{\mathcal{X}} \mathcal{K}(x,z) f(z) dP(z) = \sum_{z=1}^{d} \mathcal{K}(x,z) f(z).$$

By Linear Algebra Theory, we have

$$K = \sum_{j=1}^{d} \mu_j \boldsymbol{v}_j \boldsymbol{v}_j^\top.$$

Notations

Let $\mathbb{P}$ be a non-negative measure over a compact metric space $\mathcal{X}$, and consider the function class $L^2(\mathcal{X};\mathbb{P})$ with the usual squared norm

$$\|f\|^2_{L(\mathcal{X};\mathbb{P})} = \int_{\mathcal{X}} f^2(x) d\mathbb{P}(x).$$

Given a PSD kernel, we define a linear operator

$$T_{\mathcal{K}}(f)(x) := \int_{\mathcal{X}} \mathcal{K}(x,z) f(z) d\mathbb{P}(z).$$

We assume that

$$\int_{\mathcal{X}\times\mathcal{X}} \mathcal{K}^2(x,z) d\mathbb{P}(x) d\mathbb{P}(z) < \infty, \qquad (*)$$

which is squuared integral, then we have

$$\|T_{\mathcal{K}}(f)\|^2_{L^2(\mathcal{X})} = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \mathcal{K}(x,y) f(y) d\mathbb{P}(x) \right)^2 d\mathbb{P}(y) \leq \|f\|^2_{L(\mathcal{X})} \int_{\mathcal{X}\times\mathcal{X}} \mathcal{K}^2(x,y) d\mathbb{P}(x) d\mathbb{P}(y),$$

which implies the operator $T_{\mathcal{K}}$ is a bounded operator on $L^2(\mathcal{X})$.

Mercer's Theorem

Theorem 12.20

Suppose that $\mathcal{X}$ is compact, the kernel function $\mathcal{K}$ is continuous and positive semidefinite, and satisfies the Hilbert-Schmidt condition*. Then there exists a sequence of eigenfunctions $(\phi_j)_{j=1}^{\infty}$ that form an orthonormal basis of $L^2(\mathcal{X};\mathbb{P})$, and non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$ such that

$$T_{\mathcal{K}}(\phi_j) = \mu_j \phi_j.$$

Moreover, the kernel function has the expansion

$$\mathcal{K}(x,z) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(z),$$

where the convergence of the infinite series holds absolutely and uniformly.

The original Mercer's theorem is related to the spectral of compact operators in advanced Functional Analysis.

Examples

We define a mapping $\Phi : \mathcal{X} \to \ell^2(\mathbb{N})$ via

$$x \mapsto \Phi(x) := (\sqrt{\mu_1}\phi_1(x), \sqrt{\mu_2}\phi_2(x), \cdots).$$

By construction, we have $\|\Phi(x)\|^2_{\ell^2(\mathbb{N})} = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) = \mathcal{K}(x,x) < \infty$, which indeeed $\Phi \in \ell^2(\mathbb{N})$. Moreover,

$$\langle \Phi(x), \Phi(z) \rangle = \sum_{j=1}^{\infty} \mu_j \phi_j(x)\phi_j(z) = \mathcal{K}(x,z) < \infty,$$

thus providing a PSD kernel.

12.22 $\mathcal{K}(x,z) = (1+xz)^2$ over $[-1,1]^2$, which is equipped with Lebesgue measure. Given a function $f : [-1,1] \to \mathbb{R}$, we have

$$\int_{-1}^{1} \mathcal{K}(x,z) f(z) dz = \left(\int_{-1}^{1} f(z) dz \right) + \left(2 \int_{-1}^{1} z f(z) dz \right) x + \left(\int_{-1}^{1} z^2 f(z) dz \right) x^2.$$

So that we let the eigenfunctions be $f(x) = a_0 + a_1 x + a_2 x^2$.

Examples

We only need to solve the linear system

$$\begin{bmatrix} 2 & 0 & 2/3 \\ 0 & 4/3 & 0 \\ 2/3 & 0 & 2/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \mu \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}.$$

12.23 (Eigenfunctions for a first-order Sobolev space) The PSD takes the form $\mathcal{K}(x,z) = \min\{x,z\}$. We calculate $T_{\mathcal{K}}(\phi) = \mu\phi$, which means

$$\int_0^x z\phi(z)dz + \int_x^1 x\phi(z)dz = \mu\phi(x) \quad \forall x \in [0,1].$$

Then we take derivatives twice obtaining $\mu\phi''(x) + \phi(x) = 0$. By the definition of $\mathbb{H}^1[0,1]$, we know $\phi(0) = 0$, so that $\phi(x) = \sin(x/\sqrt{\mu})$. Taking $x = 1$ in the equation above to get $\int_0^1 z\phi(z)dz = \mu\phi(1)$, we deduce that

$$\phi_j(t) = \sin\frac{(2j-1)\pi t}{2} \quad \mu_j = \left(\frac{2}{(2j-1)\pi}\right)^2 \quad j = 1, 2, \cdots.$$

Corollary

Corollary 12.26

Consider a kernel satisfying the conditions of Mercer's Theorem with associated eigenfunctions $(\phi_j)_{j=1}^{\infty}$ and non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$. It induces the reproducing kernel Hilbert space

$$\mathcal{H} = \{f = \sum_{j=1}^{\infty} \beta_j \phi_j | \text{for some } (\beta_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N}) \text{ with } \sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} < \infty\},$$

along with inner product

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle \langle g, \phi_j \rangle}{\mu_j},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $L^2(\mathcal{X}; \mathbb{P})$.

Proof

RMK: This Cor shows that the RKHS associated with a Mercer kernel is isomorphic to an infinite-dimensional ellipsoid contained with $\ell^2(\mathbb{N})$ - namely

$$\mathcal{E} = \left\{ (\beta_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N}) | \sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} \leq 1 \right\}.$$

Proof:

We only need to verify that $\mathcal{H}$ has the reproducing property with respect to the given kernel. By Mercer's theorem, we have $\mathcal{K}(\cdot, x) = \sum_{j=1}^{\infty} \mu_j \phi_j(\cdot)\phi_j(x)$, so that $\beta_j = \mu_j \phi_j(x)$. Moreover,

$$\sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} = \sum_{j=1}^{\infty} \mu_j \phi_j^2(x) = \mathcal{K}(x, x) < \infty,$$

so that $\mathcal{K}(\cdot, x) \in \mathcal{H}$.

Proof

Let us now verify the reproducing property. By the orthonormaility of ϕ_j, we have $\langle \mathcal{K}(\cdot, x), \phi_j \rangle = \mu_j \phi_j(x)$. Thus, for any $f \in \mathcal{H}$, we have

$$\langle f, \mathcal{K}(\cdot, x) \rangle = \sum_{j=1}^{\infty} \frac{\langle f, \phi_j \rangle \langle \mathcal{K}(\cdot, x), \phi_j \rangle}{\mu_j} = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j(x) = f(x),$$

where the last equality is by the orthonormality of $(\phi_j)_{j=1}^{\infty}$.

Overview

Outline

Sums of reproducing kernels

Given two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ of functions defined on domains $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, consider the space

$$\mathbb{H}_1 + \mathbb{H}_2 := \{f_1 + f_2 \mid f_j \in \mathbb{H}_j, j = 1, 2\},$$

corresponding to the set of all functions obtained as sums of pairs of functions from the two spaces.

Proposition 12.27

Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are both *RKHS*s with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the space $\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2$ with norm

$$\|f\|_{\mathbb{H}}^2 := \min_{\substack{f=f_1+f_2 \\ f_1 \in \mathbb{H}_1, f_2 \in \mathbb{H}_2}} \left\{ \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \right\}$$

is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$.

Proof of Proposition 12.27

- Consider the direct sum $\mathbb{F} := \mathbb{H}_1 \oplus \mathbb{H}_2$ of the two Hilbert spaces; it is the Hilbert space $\{(f_1, f_2) \mid f_j \in \mathbb{H}_j, j = 1, 2\}$ of all ordered pairs, and

$$\|(f_1, f_2)\|_{\mathbb{F}}^2 := \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 .$$

$$\langle (f_1, f_2), (g_1, g_2) \rangle_{\mathbb{F}} = \langle f_1, g_1 \rangle_{\mathbb{H}_1} + \langle f_2, g_2 \rangle_{\mathbb{H}_2}.$$

  The linear operator $L : \mathbb{F} \to \mathbb{H}$ defined by $(f_1, f_2) \mapsto f_1 + f_2$.
- The nullspace $\mathbb{N}(L)$ of this operator is a subspace of $\mathbb{F}$, and we claim that it is closed. Consider some sequence $((f_n, -f_n))_{n=1}^{\infty}$ contained within the nullspace $\mathbb{N}(L)$ that converges to a point $(f, g) \in \mathbb{F}$. By the definition of the norm, this convergence implies that $f_n \to f$ in $\mathbb{H}_1$ (and hence pointwise) and $-f_n \to g$ in $\mathbb{H}_2$ (and hence pointwise). Overall, we conclude that $f = -g$, meaning $(f, g) \in \mathbb{N}(L)$.

Proof of Proposition 12.27

- Verifying the space $\mathbb{H}$ with the inner product is a Hilbert space.
  Let $\mathbb{N}^{\perp}$ be the orthogonal complement of $\mathbb{N}(L)$ in $\mathbb{F}$, and let $L_{\perp} : \mathbb{N}^{\perp}(L) \to \mathbb{H}$ be the restriction of L to $\mathbb{N}^{\perp}$. Since this map is a bijection between $\mathbb{N}^{\perp}$ and $\mathbb{H}$, we may define an inner product on $\mathbb{H}$ via

$$\langle f, g \rangle_{\mathbb{H}} := \left\langle L_{\perp}^{-1}(f), L_{\perp}^{-1}(g) \right\rangle_{\mathbb{F}}.$$

- Check that $\mathbb{H}$ is an RKHS with kernel $\mathcal{K} = \mathcal{K}_1 + \mathcal{K}_2$, and that the norm $\|\cdot\|_{\mathbb{H}}^2$ takes the given form (12.19).

Proof of Proposition 12.27

- Since the functions $\mathcal{K}_1(\cdot, x)$ and $\mathcal{K}_2(\cdot, x)$ belong to $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively, the function $\mathcal{K}(\cdot, x) = \mathcal{K}_1(\cdot, x) + \mathcal{K}_2(\cdot, x)$ belongs to $\mathbb{H}$. For a fixed $f \in \mathbb{F}$, let $(f_1, f_2) = L_\perp^{-1}(f) \in \mathbb{F}$, and for a fixed $x \in \mathcal{X}$, let $(g_1, g_2) = L_\perp^{-1}(\mathcal{K}(\cdot, x)) \in \mathbb{F}$. Since $(g_1 - \mathcal{K}_1(\cdot, x), g_2 - \mathcal{K}_2(\cdot, x))$ must belong to $\mathbb{N}(L)$, it must be orthogonal (in $\mathbb{F}$) to the element $(f_1, f_2) \in \mathbb{N}^\perp$. Consequently, we have $\langle (g_1 - \mathcal{K}_1(\cdot, x), g_2 - \mathcal{K}_2(\cdot, x)), (f_1, f_2) \rangle_{\mathbb{F}} = 0$, and hence

$$\begin{aligned} \langle f_1, \mathcal{K}_1(\cdot, x) \rangle_{\mathbb{H}_1} + \langle f_2, \mathcal{K}_2(\cdot, x) \rangle_{\mathbb{H}_2} &= \langle f_1, g_1 \rangle_{\mathbb{H}_1} + \langle f_2, g_2 \rangle_{\mathbb{H}_2} \\ &= \langle f, \mathcal{K}(\cdot, x) \rangle_{\mathbb{H}}. \end{aligned}$$

Since $\langle f_1, \mathcal{K}_1(\cdot, x) \rangle_{\mathbb{H}_1} + \langle f_2, \mathcal{K}_2(\cdot, x) \rangle_{\mathbb{H}_2} = f_1(x) + f_2(x) = f(x)$, we have established that $\mathcal{K}$ has the reproducing property.

Proof of Proposition 12.27

- Let us verify that the norm $\|f\|_{\mathbb{H}} := \left\|L_{\perp}^{-1}(f)\right\|_{\mathbb{F}}$ that we have defined is equivalent to the definition (12.19). For a given $f \in \mathbb{H}$, consider some pair $(f_1, f_2) \in \mathbb{F}$ such that $f = f_1 + f_2$, and define $(v_1, v_2) = (f_1, f_2) - L_{\perp}^{-1}(f)$. We have

$$\|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2 \overset{\text{(i)}}{=} \|(f_1, f_2)\|_{\mathbb{F}}^2 \overset{\text{(ii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \left\|L_{\perp}^{-1}(f)\right\|_{\mathbb{F}}^2$$
$$\overset{\text{(iii)}}{=} \|(v_1, v_2)\|_{\mathbb{F}}^2 + \|f\|_{\mathbb{H}}^2,$$

where step (i) uses the definition (12.22) of the norm in $\mathbb{F}$, step (ii) follows from the Pythagorean property, as applied to the pair $(v_1, v_2) \in \mathbb{N}(L)$ and $L_{\perp}^{-1}(f) \in \mathbb{N}^{\perp}$, and step (iii) uses our definition of the norm $\|f\|_{\mathbb{H}}$. Consequently, we have shown that for any pair f_1, f_2 such that $f = f_1 + f_2$, we have

$$\|f\|_{\mathbb{H}}^2 \leq \|f_1\|_{\mathbb{H}_1}^2 + \|f_2\|_{\mathbb{H}_2}^2,$$

with equality holding if and only if $(v_1, v_2) = (0, 0)$, or equivalently $(f_1, f_2) = L_{\perp}^{-1}(f)$. This establishes the equivalence of the definitions.

First-order Sobolev space and constant functions

Consider the kernel functions on $[0,1] \times [0,1]$ given by $\mathcal{K}_1(x,z) = 1$ and $\mathcal{K}_2(x,z) = \min\{x,z\}$. They generate the reproducing kernel Hilbert spaces

$$\mathbb{H}_1 = \text{span}\{1\} \quad \text{and} \quad \mathbb{H}_2 = \mathbb{H}^1[0,1],$$

$$\mathbb{H}^1[0,1] = \{f : [0,1] \to \mathbb{R} \mid f(0) = 0, \text{ and } f \text{ is abs. cts. with } f' \in L^2[0,1]\}.$$

where $\text{span}\{1\}$ is the set of all constant functions, and $\mathbb{H}^1[0,1]$ is the first-order Sobolev space.

- Note that $\mathbb{H}_1 \cap \mathbb{H}_2 = \{0\}$, since $f(0) = 0$ for any element of $\mathbb{H}_2$.
- The RKHS with kernel $\mathcal{K}(x,z) = 1 + \min\{x,z\}$ consists of all functions

  $$\bar{\mathbb{H}}^1[0,1] := \{f : [0,1] \to \mathbb{R} \mid f \text{ is absolutely continuous with } f' \in L^2[0,1]\},$$

  equipped with the squared norm $\|f\|^2_{\bar{\mathbb{H}}^1[0,1]} = f^2(0) + \int_0^1 (f'(z))^2 \, dz$.

Higher-order Sobolev spaces and polynomial space

For an integer $\alpha \geq 1$, consider the kernel functions on $[0,1] \times [0,1]$ given by

$$\mathcal{K}_1(x,z) = \sum_{\ell=0}^{\alpha-1} \frac{x^\ell}{\ell!} \frac{z^\ell}{\ell!} \quad \text{and} \quad \mathcal{K}_2(x,z) = \int_0^1 \frac{(x-y)_+^{\alpha-1}}{(\alpha-1)!} \frac{(z-y)_+^{\alpha-1}}{(\alpha-1)!} dy.$$

The first kernel generates an RKHS $\mathbb{H}_1$ of polynomials of degree $\alpha - 1$, the second kernel generates the α-order Sobolev space $\mathbb{H}_2 = \mathbb{H}^\alpha[0,1]$.

Higher-order Sobolev spaces and polynomial space

- Any function $f \in \mathbb{H}^\alpha[0,1]$ satisfies $f^{(\ell)}(0) = 0$ for $\ell = 0, 1, \ldots, \alpha - 1$, but it is not hold in $\mathbb{H}_1$. Consequently, $\mathbb{H}_1 \cap \mathbb{H}_2 = \{0\}$ holds.
- The sum of kernel

$$\mathcal{K}(x,z) = \sum_{\ell=0}^{\alpha-1} \frac{x^\ell}{\ell!}\frac{z^\ell}{\ell!} + \int_0^1 \frac{(x-y)_+^{\alpha-1}}{(\alpha-1)!}\frac{(z-y)_+^{\alpha-1}}{(\alpha-1)!}dy.$$

- The Hilbert space $\mathbb{H}^\alpha[0,1]$ of all functions that are α-times differentiable almost everywhere, with $f^{(\alpha)}$ Lebesgue-integrable.
- As we verify in Exercise 12.15, the RKHS norm takes the form

$$\|f\|_{\mathbb{H}}^2 = \sum_{\ell=0}^{\alpha-1} \left(f^{(\ell)}(0)\right)^2 + \int_0^1 \left(f^{(\alpha)}(z)\right)^2 dz.$$

everywhere, with $f^{(\alpha)}$ Lebesgue-integrable.

Exercise 12.15

For $\mathbb{H}_1$, we show that the inner product is given by

$$\langle f, g\rangle_{\mathbb{H}_1} = \sum_{\ell=0}^{\alpha-1} f^{(\ell)}(0) g^{(\ell)}(0)$$

then the norm of $\mathbb{H}_1$ is given by

$$\|f\|_{\mathbb{H}_1} = \langle f, f\rangle_{\mathbb{H}_1} = \sum_{\ell=0}^{\alpha-1} \left(f^{(\ell)}(0)\right)^2.$$

Consider a polynomial of degree $\alpha - 1$: $f(x) = \sum_{\ell=0}^{\alpha-1} a_i x^\ell$. Then,

$$f(x) = \sum_{\ell=0}^{\alpha-1} \frac{x^\ell}{\ell!} f^{(\ell)}(0) \overset{(*)}{=} \langle \mathcal{K}_1(x, \cdot), f\rangle_{\mathbb{H}_1}$$

where equality $(*)$ follows from the observation that

$$\left(\frac{\mathrm{d}}{\mathrm{d}z}\right)^\ell \mathcal{K}_1(x, z)\Big|_{z=0} = \frac{x^\ell}{\ell!}.$$

Additive models

- $\mathbb{H}_j$ be a reproducing kernel Hilbert space, $f : \mathbb{R}^d \to \mathbb{R}$, where $f_j \in \mathbb{H}_j$, to overcome the curse of dimensionality, consider the additive decomposition

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} f_j(x_j),$$

  where $f_j : \mathbb{R} \to \mathbb{R}$ is a univariate function for the j th coordinate.
- Since $\mathbb{H}_j \cap \mathbb{H}_k = \{0\}$ for all $j \neq k$, the associated Hilbert norm takes the form $\|f\|_{\mathbb{H}}^2 = \sum_{j=1}^{d} \|f_j\|_{\mathbb{H}_j}^2$.
- When the expansion functions are chosen to be mutually orthogonal, functional ANOVA decompositions as following

$$f(x_1, \ldots, x_d) = \sum_{j=1}^{d} f_j(x_j) + \sum_{j \neq k} f_{jk}(x_j, x_k) + \cdots$$

Outline

Tensor products

Consider two separable Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$ of functions, say with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. They can be used to define a new Hilbert space, denoted by $\mathbb{H}_1 \otimes \mathbb{H}_2$, known as the tensor product of $\mathbb{H}_1$ and $\mathbb{H}_2$. Consider the set of functions $h : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ that have the form

$$\{h = \sum_{j=1}^{n} f_j g_j \mid \text{for some } \ n \in \mathbb{N} \text{ and such that } f_j \in \mathbb{H}_1, g_j \in \mathbb{H}_2 \text{ for all } j \in [n]\}.$$

If $h = \sum_{j=1}^{n} f_j g_j$ and $\widetilde{h} = \sum_{k=1}^{m} \widetilde{f}_k \widetilde{g}_k$ are two members of this set, we define their inner product

$$\langle h, \widetilde{h} \rangle_{\mathbb{H}} := \sum_{j=1}^{n} \sum_{k=1}^{m} \left\langle f_j, \widetilde{f}_k \right\rangle_{\mathbb{H}_1} \langle g_j, \widetilde{g}_k \rangle_{\mathbb{H}_2} .$$

Proposition 12.31

Suppose that $\mathbb{H}_1$ and $\mathbb{H}_2$ are reproducing kernel Hilbert spaces of real-valued functions with domains $\mathcal{X}_1$ and $\mathcal{X}_2$, and equipped with kernels $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. Then the tensor product space $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$ is an RKHS of real-valued functions with domain $\mathcal{X}_1 \times \mathcal{X}_2$, and with kernel function

$$\mathcal{K}\left((x_1, x_2), (x_1', x_2')\right) = \mathcal{K}_1\left(x_1, x_1'\right)\mathcal{K}_2\left(x_2, x_2'\right).$$

Proof of Proposition 12.31

The Schur product theorem states that the Hadamard product of two positive definite matrices is also a positive definite matrix, then $\mathcal{K}$ is a positive semidefinite function.
By definition of the tensor product space $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$, for each pair $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, the function $\mathcal{K}((\cdot,\cdot),(x_1,x_2)) = \mathcal{K}_1(\cdot, x_1)\mathcal{K}_2(\cdot, x_2)$ is an element of the tensor product space $\mathbb{H}$. Let $f = \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j\psi_k$ be an arbitrary element of $\mathbb{H}$. We have

$$\begin{aligned}\langle f, \mathcal{K}((\cdot,\cdot),(x_1,x_2))\rangle_{\mathbb{H}} &= \sum_{j,k=1}^{n} \alpha_{j,k} \langle \phi_j, \mathcal{K}_1(\cdot, x_1)\rangle_{\mathbb{H}_1} \langle \psi_k, \mathcal{K}_2(\cdot, x_2)\rangle_{\mathbb{H}_2} \\ &= \sum_{j,k=1}^{n} \alpha_{j,k}\phi_j(x_1)\psi_k(x_2) = f(x_1, x_2),\end{aligned}$$

thereby verifying the reproducing property.

Overview

Outline

Function interpolation by kernel

Suppose that we observe n samples of an unknown function $f^* : \mathcal{X} \to \mathbb{R}$, say of the form $y_i = f^*(x_i)$ for $i = 1, 2, \ldots, n$, where the design sequence $\{x_i\}_{i=1}^n$ is known to us.

By minimizing RKHS norm, it can be formulated as an optimization problem Hilbert space,

$$\text{choose } \widehat{f} \in \arg\min_{f \in \mathbb{H}} \|f\|_{\mathbb{H}} \quad \text{such that} \quad f(x_i) = y_i \text{ for } i = 1, 2, \ldots, n. \tag{1}$$

Function interpolation

Proposition 12.32

Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix defined by the design points $\{x_i\}_{i=1}^n$. The convex program (1) is feasible if and only if $y \in$ range $(\mathbf{K})$, in which case any optimal solution can be written as

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \mathcal{K}(\cdot, x_i), \qquad \text{where } \mathbf{K}\widehat{\alpha} = y/\sqrt{n}.$$

- The kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$.

Proof of Proposition 12.32

For a given vector $\alpha \in \mathbb{R}^n$, define the function $f_\alpha(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \mathcal{K}(\cdot, x_i)$, and consider the set $\mathbb{L} := \{f_\alpha \mid \alpha \in \mathbb{R}^n\}$. Note that for any $f_\alpha \in \mathbb{L}$, we have

$$f_\alpha(x_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \mathcal{K}(x_j, x_i) = \sqrt{n}(\mathbf{K}\alpha)_j,$$

where $(\mathbf{K}\alpha)_j$ is the j th component of the vector $\mathbf{K}\alpha \in \mathbb{R}^n$.

- The function $f_\alpha \in \mathbb{L}$ satisfies the interpolation condition if and only if $\mathbf{K}\alpha = y/\sqrt{n}$.
- "$\Leftarrow$" If $y \in \text{range}(\mathbf{K})$, $f_\alpha \in \mathbb{L}$ satisfies the interpolation condition.
- "$\Rightarrow$" Note that $\mathbb{L}$ is a finite-dimensional (hence closed) linear subspace of $\mathbb{H}$. Consequently, any function $f \in \mathbb{H}$ can be decomposed uniquely as $f = f_\alpha + f_\perp$, where $f_\alpha \in \mathbb{L}$ and $f_\perp$ is orthogonal to $\mathbb{L}$. Using this decomposition and the reproducing property, we have

$$y_j = f(x_j) = \langle f, \mathcal{K}(\cdot, x_j)\rangle_{\mathbb{H}} = \langle f_\alpha + f_\perp, \mathcal{K}(\cdot, x_j)\rangle_{\mathbb{H}} = f_\alpha(x_j),$$

where $\mathcal{K}(\cdot, x_j) \in \mathbb{L}$, and $\|f_\alpha + f_\perp\|_{\mathbb{H}}^2 = \|f_\alpha\|_{\mathbb{H}}^2 + \|f_\perp\|_{\mathbb{H}}^2$, deduce $f_\perp = 0$.

Advantages

The third optimality result is in the context of quasi-interpolation, i.e.,

Theorem (Optimality III)

Suppose $K \in C(\Omega \times \Omega)$ is a strictly positive definite kernel and suppose that $\boldsymbol{x} \in \Omega$ is fixed. Let $\hat{u}_j(\boldsymbol{x}), j = 1, \ldots, N$, be the values at $\boldsymbol{x}$ of the cardinal basis functions for interpolation with K. Then

$$\left| f(\boldsymbol{x}) - \sum_{j=1}^{N} f(\boldsymbol{x}_j)\, u_j^\star(\boldsymbol{x}) \right| \leq \left| f(\boldsymbol{x}) - \sum_{j=1}^{N} f(\boldsymbol{x}_j)\, u_j \right|$$

for all choices of $u_1, \ldots, u_N \in \mathbb{R}$.

- The cardinal form of the kernel interpolant is **more accurate** (as measured by the pointwise error) than any other linear combination of the data.

Experiments

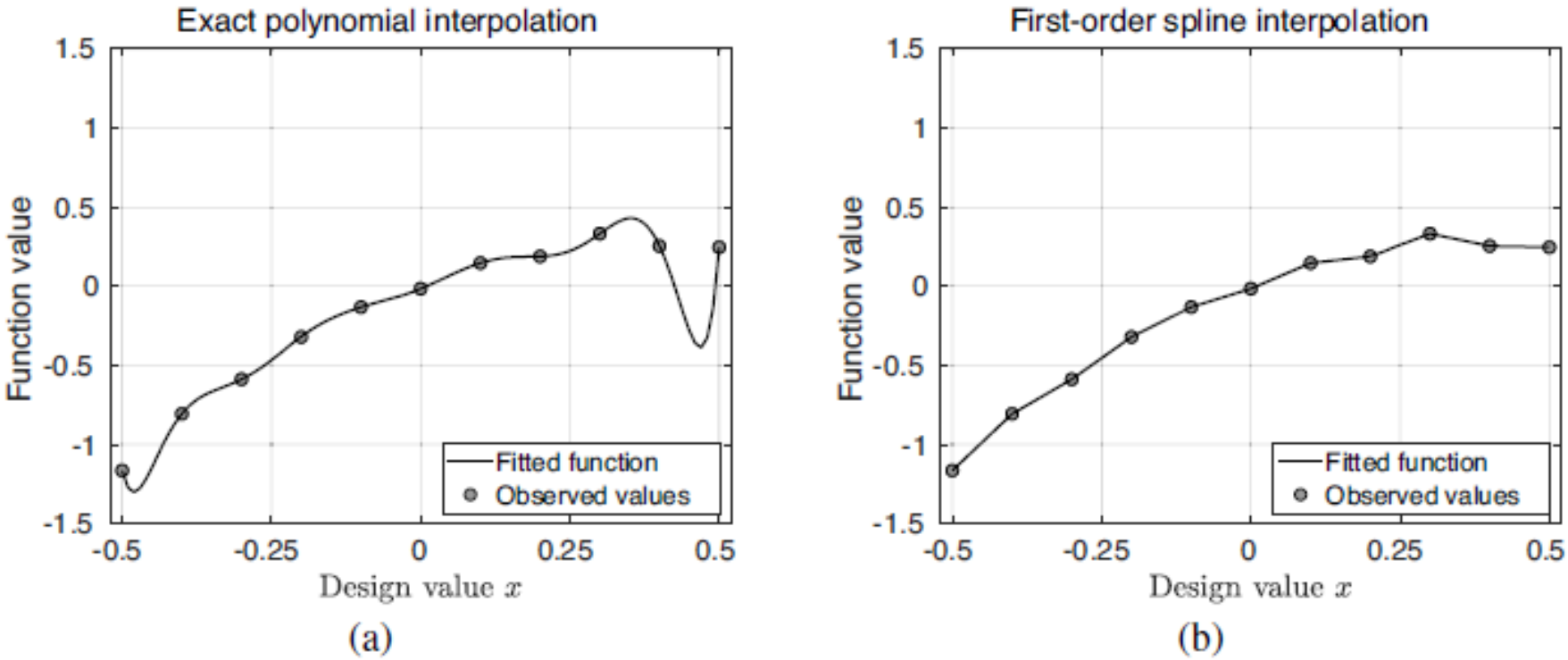

Figure 12.1 Exact interpolation of $n = 11$ equally sampled function values using RKHS methods. (a) Polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^{12}$. (b) First-order Sobolev kernel $\mathcal{K}(x, z) = 1 + \min\{x, z\}$.

Outline

Fitting via kernel ridge regression

- Consider a noisy observation model

$$y_i = f^*(x_i) + w_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where the coefficients $\{w_i\}_{i=1}^n$ model noisiness or disturbance in the measurement model.

- Transforming the model to the optimization problem

$$\min_{f \in \mathbb{H}} \|f\|_{\mathbb{H}} \quad \text{such that } \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \delta^2,$$

where $\delta > 0$ is some type of tolerance parameter.

- Alternatively, we might minimize the mean-squared error subject to a bound on the Hilbert radius of the solution, say

$$\min_{f \in \mathbb{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad \text{such that } \|f\|_{\mathbb{H}} \leq R.$$

for an appropriately chosen radius $R > 0$.

Fitting via kernel ridge regression (KRR)

- Both of these problems are convex, and so by Lagrangian duality, they can be reformulated in the penalized form

$$\widehat{f} = \arg\min_{f \in \mathbb{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \right\}. \tag{2}$$

  Here, for a fixed set of observations $\{(x_i, y_i)\}_{i=1}^n$, the regularization parameter $\lambda_n \geq 0$ is a function of the tolerance δ or radius R.

Fitting via kernel ridge regression (KRR)

Consider the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$.

Proposition 12.33

For all $\lambda_n > 0$, the kernel ridge regression estimate (2) can be written as

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \mathcal{K}(\cdot, x_i), \tag{3}$$

where the optimal weight vector $\widehat{\alpha} \in \mathbb{R}^n$ is given by

$$\widehat{\alpha} = (\mathbf{K} + \lambda_n \mathbf{I}_n)^{-1} \frac{y}{\sqrt{n}}. \tag{4}$$

Proof of Proposition 12.33

- Similar as the proof of Proposition 12.32, any optimal solution must be expressed as the sum of kernel (3).
- It remains to prove the specific form (4) of the optimal $\widehat{\alpha}$.
  Given a function f of the form (3), for each $j = 1, 2, \ldots, n$, we have

$$f(x_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i \mathcal{K}(x_j, x_i) = \sqrt{n} e_j^{\mathrm{T}} \mathbf{K} \alpha,$$

  where $e_j \in \mathbb{R}^n$ is the canonical basis vector with 1 in position j, and we have recalled that $K_{ji} = \mathcal{K}(x_j, x_i)/n$. Then,

$$\|f\|_{\mathbb{H}}^2 = \frac{1}{n} \left\langle \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x_i), \sum_{j=1}^{n} \alpha_j \mathcal{K}(\cdot, x_j) \right\rangle_{\mathbb{H}} = \alpha^{\mathrm{T}} \mathbf{K} \alpha.$$

Proof of Proposition 12.33

- Substituting these relations into the cost function, we have

$$\frac{1}{n}\|y-\sqrt{n}\mathbf{K}\alpha\|_2^2+\lambda\alpha^{\mathrm{T}}\mathbf{K}\alpha=\frac{1}{n}\|y\|_2^2+\alpha^{\mathrm{T}}\left(\mathbf{K}^2+\lambda\mathbf{K}\right)\alpha-\frac{2}{\sqrt{n}}y^{\mathrm{T}}\mathbf{K}\alpha.$$

In order to find the minimum of this quadratic function, we compute the gradient of α and set it equal to zero, thereby

$$\mathbf{K}\left(\mathbf{K}+\lambda\mathbf{I}_n\right)\alpha=\mathbf{K}\frac{y}{\sqrt{n}}.$$

Thus, we see that the vector $\widehat{\alpha}$ previously defined in equation (4) is optimal.

Experiments

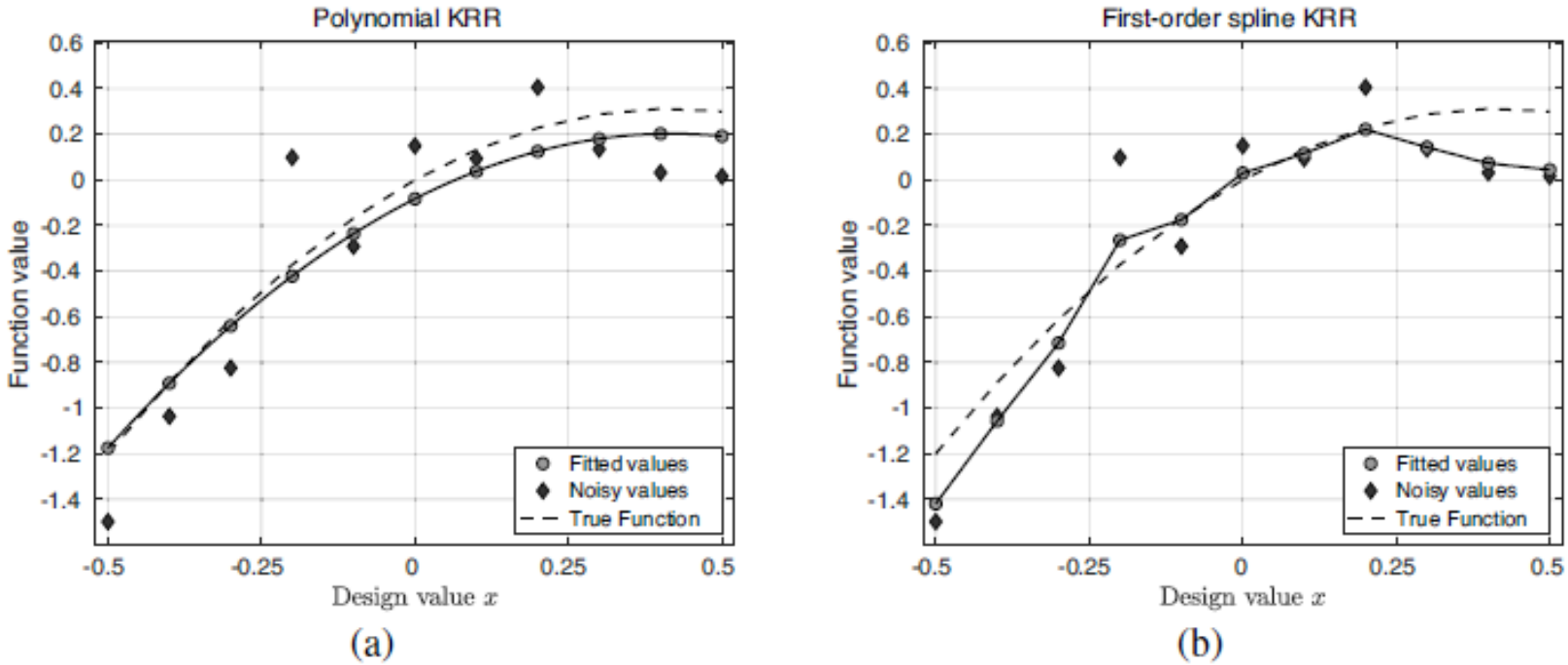

Figure 12.2 Illustration of kernel ridge regression estimates of function $f^*(x) = \frac{3x}{2} - \frac{9}{5}x^2$ based on $n = 11$ samples, located at design points $x_i = -0.5 + 0.10\,(i-1)$ over the interval $[-0.5, 0.5]$. (a) Kernel ridge regression estimate using the second-order polynomial kernel $\mathcal{K}(x, z) = (1 + xz)^2$ and regularization parameter $\lambda_n = 0.10$. (b) Kernel ridge regression estimate using the first-order Sobolev kernel $\mathcal{K}(x, z) = 1 + \min\{x, z\}$ and regularization parameter $\lambda_n = 0.10$.

Overview

Measure mean discrepancies

Let $\mathscr{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ that are integrable with respect to $\mathbb{P}$ and $\mathbb{Q}$. We can then define the quantity

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathscr{F}} \left| \int f(d\mathbb{P} - d\mathbb{Q}) \right| = \sup_{f \in \mathscr{F}} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)] \right|.$$

- For any choice of function class $\mathscr{F}$, this always defines a pseudometric, meaning that $\rho_{\mathscr{F}}$ satisfies all the metric properties, except that there may exist pairs $\mathbb{P} \neq \mathbb{Q}$ such that $\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = 0$.
- When $\mathscr{F}$ is sufficiently rich, then $\rho_{\mathscr{F}}$ becomes a metric, known as an integral probability metric.

Kolmogorov metric (non-RKHS function classes)

Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are measures on the real line. For each $t \in \mathbb{R}$, let $\mathbb{I}_{(-\infty,t]}$ denote the $\{0,1\}$-valued indicator function for the event $\{x \leq t\}$, and consider the function class $\mathscr{F} = \left\{\mathbb{I}_{(-\infty,t]} \mid t \in \mathbb{R}\right\}$. We then have

$$\rho_{\mathscr{F}}(\mathbb{P},\mathbb{Q}) = \sup_{t\in\mathbb{R}} |\mathbb{P}(X \leq t) - \mathbb{Q}(X \leq t)| = \|F_{\mathbb{P}} - F_{\mathbb{Q}}\|_{\infty},$$

where $F_{\mathbb{P}}$ and $F_{\mathbb{Q}}$ are the cumulative distribution functions of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Thus, this choice leads to the Kolmogorov distance between $\mathbb{P}$ and $\mathbb{Q}$.

Total variation distance (non-RKHS function classes)

Consider the class $\mathscr{F} = \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_\infty \leq 1\}$ of real-valued functions bounded by one in the supremum norm. With this choice, we have

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_\infty \leq 1} \left| \int f(d\mathbb{P} - d\mathbb{Q}) \right|.$$

In Exercise 3.13, we established that

$$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{f:\mathcal{X}\to[0,1]} \int f(p - q)\mathrm{d}\mu.$$

Therefore,

$$2\,\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{f:\mathcal{X}\to[-1,1]} \int f(p - q)\mathrm{d}\mu = \sup_{\|f\|_\infty \leq 1} \int f(p - q)\mathrm{d}\mu,$$

this metric corresponds to (two times) the total variation distance

$$\rho_{\mathscr{F}}(\mathbb{P}, \mathbb{Q}) = 2 \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where the supremum ranges over all measurable subsets of $\mathcal{X}$.

Kernel means discrepancy (KMD)

Given an RKHS with kernel function $\mathcal{K}$, consider the pseudometric

$$\rho_{\mathbb{H}}(\mathbb{P},\mathbb{Q}) := \sup_{\|f\|_{\mathbb{H}}\leq 1} \left|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)]\right|.$$

Define a mean embedding $\mu_{\mathbb{P}}(t) = \langle \mu_{\mathbb{P}}, \mathcal{K}(t,\cdot)\rangle_{\mathbb{H}} = \mathbb{E}_x \mathcal{K}(t,x)$.

$$\begin{aligned}\rho^2_{\mathbb{H}}(\mathbb{P},\mathbb{Q}) &= \left(\sup_{\|f\|_{\mathbb{H}}\leq 1} \left|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Z)]\right|\right)^2 \\ &= \left(\sup_{\|f\|_{\mathbb{H}}\leq 1} \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, f\rangle_{\mathbb{H}}\right)^2 = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2_{\mathbb{H}}\end{aligned} \tag{5}$$

Then this pseudometric as a kernel means discrepancy (KMD),

$$\rho^2_{\mathbb{H}}(\mathbb{P},\mathbb{Q}) = \mathbb{E}\left[\mathcal{K}\left(X,X'\right) + \mathcal{K}\left(Z,Z'\right) - 2\mathcal{K}(X,Z)\right],$$

where $X, X' \sim \mathbb{P}$ and $Z, Z' \sim \mathbb{Q}$ are all mutually independent random vectors.

KMD for linear kernel

Let us compute the KMD for the linear kernel $\mathcal{K}(x,z) = \langle x, z\rangle$ on $\mathbb{R}^d$. Letting $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on $\mathbb{R}^d$ with mean vectors $\mu_p = \mathbb{E}_{\mathbb{P}}[X]$ and $\mu_q = \mathbb{E}_{\mathbb{Q}}[Z]$, respectively, we have

$$\begin{aligned}\rho^2_{\mathbb{H}}(\mathbb{P},\mathbb{Q}) &= \mathbb{E}\left[\left\langle X, X'\right\rangle + \left\langle Z, Z'\right\rangle - 2\langle X, Z\rangle\right] \\ &= \|\mu_p\|_2^2 + \|\mu_q\|_2^2 - 2\left\langle \mu_p, \mu_q\right\rangle \\ &= \|\mu_p - \mu_q\|_2^2 .\end{aligned}$$

The KMD is a pseudometric for the linear kernel.

KMD for polynomial kernel

Let us consider the homogeneous polynomial kernel of degree two, namely $\mathcal{K}(x,z) = \langle x,z\rangle^2$. For this choice of kernel, we have

$$\mathbb{E}\left[\mathcal{K}\left(X,X'\right)\right] = \mathbb{E}\left[\left(\sum_{j=1}^{d} X_j X_j'\right)^2\right] = \sum_{i,j=1}^{d} \mathbb{E}\left[X_i X_j\right] \mathbb{E}\left[X_i' X_j'\right] = \|\Gamma_p\|_F^2,$$

where $\Gamma_p \in \mathbb{R}^{d\times d}$ is the second-order moment matrix with entries $[\Gamma_p]_{ij} = \mathbb{E}\left[X_i X_j\right]$. Similarly, we have $\mathbb{E}\left[\mathcal{K}\left(Z,Z'\right)\right] = \|\Gamma_q\|_F^2$, where Γ_q is the second-order moment matrix for $\mathbb{Q}$. Finally, similar calculations yield that

$$\mathbb{E}[\mathcal{K}(X,Z)] = \sum_{i,j=1}^{d} [\Gamma_p]_{ij}\,[\Gamma_q]_{ij} = \langle\langle \Gamma_p, \Gamma_q\rangle\rangle$$

where $\langle\langle\cdot,\cdot\rangle\rangle$ denotes the trace inner product between symmetric matrices. Putting together the pieces, we conclude that

$$\rho_{\mathbb{H}}^2(\mathbb{P},\mathbb{Q}) = \|\Gamma_p - \Gamma_q\|_F^2.$$

KMD for a first-order Sobolev kernel

Let us now consider the KMD induced by the kernel function $\mathcal{K}(x,z) = \min\{x,z\}$, defined on the Cartesian product $[0,1] \times [0,1]$. the kernel function generates the first-order Sobolev space

$$\mathbb{H}^1[0,1] = \left\{ f : \mathbb{R}[0,1] \to \mathbb{R} \mid f(0) = 0 \text{ and } \int_0^1 \left(f'(x)\right)^2 dx < \infty \right\},$$

with Hilbert norm $\|f\|^2_{\mathbb{H}^1[0,1]} = \int_0^1 \left(f'(x)\right)^2 dx$. With this choice, we have

$$\rho^2_{\mathbb{H}}(\mathbb{P},\mathbb{Q}) = \mathbb{E}\left[\min\left\{X,X'\right\} + \min\left\{Z,Z'\right\} - 2\min\{X,Z\}\right].$$

Thank you !