Probability Error Bounds for Approximation of Functions in RKHS

Aurelian Gheondea

Institute of Mathematics of the Romanian Academy — Bilkent University

RKHS Seminar at Middle East Technical University, Ankara

Dec 10th 2024

A.D. Aydın, A. Gheondea, Reproducing kernel Hilbert spaces approximation bounds, *J. Function Spaces*, Volume 2021, Article ID 6617774, 15 pages, arXiv: https://arxiv.org/abs/2003.12801

Often, one wants to show that some random variable, or a set of random variables, is close to its expected value (mean) with a high probability. Results of this kind are called *Concentration Inequalities*.

Theorem (Hoeffding's Inequality — 1963)

Let (P, X, Ω) be a probability space, $f_1, \ldots, f_n$ independent random variables with the same expected value μ and such that $a_j \leq f_j \leq b_j$ for all $j = 1, \ldots, n$. Then, for any $\epsilon > 0$ we have

$$P\left(\left| \frac{1}{n} \sum_{j=1}^{n} f_j - \mu \right| > \epsilon \right) \leq 2 \exp\left(-\frac{2n^2\epsilon^2}{\sum_{j=1}^{n}(b_j - a_j)^2}\right).$$

WASSILY HOEFFDING, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* **58** (1963), 13–30.

Theorem (Markov-Bienaymé-Chebyshev's Inequality)

Let $(X; \Sigma; P)$ *be a probability space,* $(\mathcal{B}; \|\cdot\|)$ *be a Banach space,* $h\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ *a nondecreasing function, and let* $g\colon X \to \mathcal{B}$ *be a Borel measurable function. Then, for any* $\delta > 0$*, we have*

$$P(\{x \in X \mid \|g(x)\| \geq \delta\}) \leq \frac{1}{h(\delta)} \int_X h(\|g(x)\|) \,\mathrm{d}P(x).$$

Proof. For any $\delta > 0$ let $S_\delta := \{x \in X \mid \|g(x)\| \geq \delta\}$ and observe that, since g is Borel measurable it follows that $S_\delta \in \Sigma$. Since h is nondecreasing it is measurable and we have $h(\|g(x)\|) \geq h(\delta)$ for all $x \in S_\delta$. Then, since P is nonnegative, we have

$$\begin{aligned} \frac{1}{h(\delta)} \int_X h(\|g(x)\|) \,\mathrm{d}P(x) &\geq \frac{1}{h(\delta)} \int_{S_\delta} h(\|g(x)\|) \,\mathrm{d}P(x) \\ &\geq \frac{1}{h(\delta)} \int_{S_\delta} h(\delta) \,\mathrm{d}P(x) = P(S_\delta). \quad \square \end{aligned}$$

This inequality is mostly used when $h(t) = t^p$ for some $0 < p < \infty$. In particular, for $p = 2$, we get the following

Corollary

Let $(X; \Sigma; P)$ be a probability space, $(\mathcal{B}; \|\cdot\|)$ a Banach space, and let $g\colon X \to \mathcal{B}$ be a Borel measurable function. Then, for any $\delta > 0$, we have

$$P(\{x \in X \mid \|g(x)\| \geq \delta\}) \leq \frac{1}{\delta^2} \int_X \|g(x)\|^2 \, \mathrm{d}P(x). \tag{1.1}$$

The classical Bienaymé-Chebyshev Inequality

$$P(\{x \in X \mid |f(x) - E(f)| \geq k\sigma\}) \leq \frac{1}{k^2},$$

is obtained from the previous corollary applied for $\mathcal{B} = \mathbb{R}$, $g(x) = f(x) - E(f)$, and $\delta = k\sigma$, for $k > 0$, where $E(f) = \int_X f(x) \, \mathrm{d}x$ is the expected value of the random variable f and $\sigma^2 = E((f - E(f))^2) = E(f^2) - E(f)^2 > 0$ is the variance of f.

Uniform Bounds for Functions Approximation

The Vapnik-Chervonenkis theory in statistical learning theory relies on concentration inequalities such as Hoeffding's inequality to bound the supremum distance between expected and empirical risk.

V.N. Vapnik, *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1998.

V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., Springer, 2000.

V.N. Vapnik, A.Ya. Chervonenkis, *Theory of Pattern Recognition. Statistical Problems of Learning* [Russian], Izdat. "Nauka", Moscow, 1974.

The theory considers a *data space* $X \subseteq \mathbb{R}^m \times \mathbb{R}$ on which an unknown probability distribution P is defined, a *hypothesis set* $\mathcal{H}$ and a *loss function* $V\colon \mathcal{H} \times X \to \mathbb{R}_+$, such that one wishes to find a hypothesis $h \in \mathcal{H}$ that minimizes the *expected risk*

$$R[h] := \int_X V(h, x)\, \mathrm{d}P(x).$$

Since P is not known in general, instead of minimizing the expected risk one usually minimizes the *empirical risk*

$$\widehat{R}_S[h] = \frac{1}{n} \sum_{i=1}^{n} V(h, x_i)$$

over a finite set $S = \{x_i\}_{i=1}^n \subseteq X$ of samples. Vapnik-Chervonenkis theory measures the probability with which the maximum distance between R and $\widehat{R}$ falls below a given threshold.

Definition (Vapnik-Chervonenkis 1968, 1971, 1978)

The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{H}$ with respect to V is the maximum cardinality of finite subsets $Y \subseteq X$ that can be *shattered* by $\mathcal{H}$, i.e. for each $Y' \subseteq Y$, there exist $h \in \mathcal{H}$, $\alpha \in \mathbb{R}$ such that

$$Y' = \{x \in Y \mid V(h, x) \geq \alpha\};$$
$$Y \setminus Y' = \{x \in Y \mid V(h, x) < \alpha\}.$$

Theorem (Vapnik, Chervonenkis 1991)

Suppose $A \leq V(h,x) \leq B$ *for each* $h \in \mathcal{H}, x \in X$*, the VC dimension of* $\mathcal{H}$ *is* $d < \infty$*, and let the sampling data* $S = \{x_1, \ldots, x_n\}$ *be selected according to the probability* P*. Then, for any* $\eta \in (0,1)$*,*

$$P\left(\sup_{h\in\mathcal{H}}\left|R[h] - \widehat{R}_S[h]\right| \geq (B-A)\sqrt{\frac{d\log\frac{2en}{d} - \log\frac{\eta}{4}}{n}}\right) \leq \eta,$$

or equivalently, for each $\delta > 0$*,*

$$P\left(\sup_{h\in\mathcal{H}}\left|R[h] - \widehat{R}_S[h]\right| \geq \delta\right) \leq 4\mathrm{e}^{d\log\frac{2en}{d} - \frac{n\delta^2}{(B-A)^2}}.$$

F. GIROSI, Approximation error bounds that use VC bounds, in *Proc. International Conference on Artificial Neural Networks*, F. Fogelman-Soulied and P. Gallinari (eds), pp. 295–302, Paris 1995.

F. Girosi has used this general result to bound the uniform distance between integrals $\int J(x,y)\lambda(y)\,\mathrm{d}y$ and sums of the form $\frac{1}{n}\sum_{i=1}^{n} J(x,x_i)$, by reinterpreting $\mathcal{H}$ as $\mathbb{R}^d$, V as J and $\mathrm{d}P(y)$ as $\frac{|\lambda(y)|}{\|\lambda\|_{L^1}}\,\mathrm{d}y$:

Theorem (Girosi, [10, Proposition 2])

Let $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = \int J(x, y)\lambda(y)\,\mathrm{d}y$ *where* $\lambda \in L^1(\mathbb{R}^d)$ *and* $A \leq J(x, y) \leq B$ *for each* $x, y \in \mathbb{R}^d$. *Let* h *be the VC dimension of* $\{J(x, \cdot)\}_{x\in\mathbb{R}^d}$. *Then, with respect to the probability measure* $P(E) = \int_E \frac{|\lambda(t)|}{\|\lambda\|_{L^1}}\,\mathrm{d}t$ *on* $\mathbb{R}^d$, *for each* $\eta \in (0, 1)$ *we have*

$$P\left(\left\|f - \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(\lambda(x_i))\,\|\lambda\|_{L^1} J(\cdot, x_i)\right\|_{L^\infty} \geq 4(B - A)\|\lambda\|_{L^1}\sqrt{\frac{h\log\frac{2en}{h} - \log\frac{\eta}{4}}{n}}\right) \leq \eta,$$

where the probability P *is extended to the product probability on* $(\mathbb{R}^d)^n$, *and* $S = \{x_1, \ldots, x_n\}$ *denotes an arbitrary sample in* $\mathbb{R}^d$

M.A. Kon, L.A. Raphael, Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory, *Wavelets and Splines: Athens 2005*, pp. 271–286, *Mod. Methods Math.*, Nashboro Press, Brentwood, TN, 2006.

Kon and Raphael [10] then apply this methodology to obtain uniform approximation bounds of functions in reproducing kernel Hilbert spaces. They consider two cases where the Hilbert space is dense in $L^2(\mathbb{R}^d)$ with a stronger norm, and where it is a closed subspace with the same norm.

When $K \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ is a positive semidefinite kernel, K defines a positive compact operator on $L^2(\mathbb{R}^d)$:

$$(T_K f)(x) = \int_{\mathbb{R}^d} f(y) K(x, y) \,\mathrm{d}y$$

for $f \in L^2(\mathbb{R}^d)$. Then T_K admits a positive square root $\sqrt{T_K}$, itself induced by a kernel $\sqrt{K} \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$.

Recall that, given a Hilbert space $\mathcal{G}$ and letting $\mathcal{B}(\mathcal{G})$ denote the collection of all linear bounded operators $T\colon \mathcal{G} \to \mathcal{G}$, T is *positive* if $\langle Tg, g \rangle \geq 0$ for all $g \in \mathcal{G}$, and then there exists uniquely another positive operator $B \in \mathcal{B}(\mathcal{G})$ such that $B^2 = T$, for which we use the notation $T =: \sqrt{T} =: T^{1/2}$.

Also, T is *compact* if it can be approximated uniformly by finite rank operators.

Proposition

Let $K \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ be positive semidefinite. Then the reproducing kernel Hilbert space associated to K is $\mathcal{H}_K = \sqrt{T_K} L^2(\mathbb{R}^d)$, with $\langle f, g \rangle_{\mathcal{H}_K} = \langle \sqrt{T_K}^{-1} f, \sqrt{T_K}^{-1} g \rangle_{L^2(\mathbb{R}^d)}$ for $f, g \in \mathcal{H}_K$.

For some kernels, such as *sinc*, *wavelet* and *spline kernels*, $\mathcal{H}_K$ is a closed subspace of $L^2(\mathbb{R}^d)$ and T_K is the projection to $\mathcal{H}_K$, hence equal to its square root.

For some other kernels, such as Gaussian or Laplace kernels, $\mathcal{H}_K$ is instead dense in $L^2(\mathbb{R}^d)$ with a stronger norm.

We have the following bounds for the two cases, in which functions in the RKHS are represented as integrals in different ways.

Theorem (Kon and Raphael [10, Theorem 4])

Let $K \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ be positive semidefinite such that $\mathcal{H}_K$ is dense in $L^2(\mathbb{R}^d)$. Suppose there exist positive functions g and k such that $g \in L^2(\mathbb{R}^d)$, k is bounded away from 0, and

$$\operatorname*{ess\,sup}_{x,y \in \mathbb{R}^d} \left| \frac{\sqrt{K}(x,y)}{g(y)k(x)} \right| \leq \tau.$$

Let h be the VC dimension of $\{\frac{\sqrt{K}(x,\cdot)}{g(\cdot)k(x)}\}_{x \in \mathbb{R}^d}$. Let $f \in \mathcal{H}_K$.

Theorem (Continued)

Then with respect to the probability measure

$$\int \mathrm{d}P(x) = \int \frac{(\sqrt{T_K}^{-1} f)(x) g(x)}{\|(\sqrt{T_K}^{-1} f) g\|_{L^1}} \, \mathrm{d}x,$$

where $(\sqrt{T_K}^{-1} f) g \in L^1(\mathbb{R}^d)$ *by the Schwarz inequality since* $\sqrt{T_K}^{-1} f, g \in L^2(\mathbb{R}^d)$,

$$\begin{aligned} \frac{f(x)}{k(x)} &= \frac{1}{k(x)} \left(\sqrt{T_K} (\sqrt{T_K}^{-1} f) \right)(x) \\ &= \int \frac{\sqrt{K}(x, y)}{g(y) k(x)} (\sqrt{T_K}^{-1} f)(y) g(y) \, \mathrm{d}y, \end{aligned}$$

Theorem (Continued)

hence for any $\eta > 0$, $n \in \mathbb{N}$, setting
$c(x) = \frac{\mathrm{sgn}\left((\sqrt{T_K}^{-1} f)(x)\right)}{g(x)} \|(\sqrt{T_K}^{-1} f) g\|_{L^1}$ *we have*

$$P\left(\left\| f - \frac{1}{n}\sum_{i=1}^{n} c(x_i)\sqrt{K}(\cdot, x_i) \right\|_{L^{\infty},1/k} \geq 4\tau \|f\|_{\mathcal{H}_K} \|g\|_{L^2} \sqrt{\frac{h \log \frac{2en}{h} - \log \frac{\eta}{4}}{n}} \right) \leq \eta,$$

where $\|f\|_{L^{\infty},1/k} := \|\frac{f}{k}\|_{L^{\infty}}$.

Theorem (Kon and Raphael [10, Theorem 5])

Let $K \in L^2(\mathbb{R}^d \times \mathbb{R}^d)$ be positive semidefinite such that $\mathcal{H}_K$ is a closed subspace of $L^2(\mathbb{R}^d)$ with the L^2 norm. Suppose there exist positive functions g and k such that $g \in L^2(\mathbb{R}^d)$, k is bounded away from 0, and

$$\operatorname{ess}\sup_{x,y\in\mathbb{R}^d} \left| \frac{K(x,y)}{g(y)k(x)} \right| \leq \tau.$$

Let h be the VC dimension of $\{\frac{K(x,\cdot)}{g(\cdot)k(x)}\}_{x\in\mathbb{R}^d}$. Let $f \in \mathcal{H}_K$.

Theorem (continued)

Then with respect to the probability measure

$$\int \mathrm{d}P(x) = \int \frac{f(x)g(x)}{\|fg\|_{L^1}} \,\mathrm{d}x,$$

where $fg \in L^1(\mathbb{R}^d)$ *by the Schwarz inequality since* $f, g \in L^2(\mathbb{R}^d)$,

$$\frac{f(x)}{k(x)} = \frac{1}{k(x)}(T_K f)(x) = \int \frac{K(x,y)}{g(y)k(x)} f(y)g(y)\,\mathrm{d}y,$$

hence for any $\eta > 0$, $n \in \mathbb{N}$, *setting* $c(x) = \frac{\operatorname{sgn}(f(x))}{g(x)}\|fg\|_{L^1}$ *we have*

$$P\left(\left\| f - \frac{1}{n}\sum_{i=1}^{n} c(x_i)K(\cdot, x_i)\right\|_{L^\infty, 1/k} \geq 4\tau\|f\|_{L^2}\|g\|_{L^2}\sqrt{\frac{h\log\frac{2en}{h} - \log\frac{\eta}{4}}{n}}\right) \leq \eta.$$

While these bounds guarantee uniform convergence in probability, the approximating functions are not orthogonal projections of f nor necessarily elements of a reproducing kernel Hilbert space, and hence may not capture f exactly at $(x_i)_{i=1}^n$ nor converge monotonically. Furthermore, the fact that the norm is not a RKHS norm means that derivatives of f may not be approximated in general, since differentiation is not bounded with respect to the uniform norm.

The Regularised Multiview Learning Problem

Let X be a nonempty set and $\mathbf{W} = \{\mathcal{W}_x\}_{x \in X}$ be a bundle of Hilbert spaces on X. In this section, it is not important whether the Hilbert spaces are complex or real, hence all Hilbert spaces are considered to be over the field $\mathbb{F}$, that is either $\mathbb{C}$ or $\mathbb{R}$. There is a difference between the complex and the real case consisting in the fact that in the latter case, for positive semidefiniteness we assume also the symmetry, or Hermitian, property, while in the complex case, the symmetry property is a consequence of the positive semidefiniteness. If K is a positive semidefinite $\mathbf{W}$-operator valued kernel, we let $\mathcal{H}_K$ be its reproducing kernel Hilbert space, as in the previous subsection.

Also, let $\mathrm{Y} = \{\mathcal{Y}_x\}_{x \in X}$ be a bundle of Hilbert spaces.
For $l, u \in \mathbb{N}$, consider input distinct points $x_1, \ldots, x_{l+u} \in X$. Here $x_1, \ldots, x_l$ are the labeled input points while $x_{l+1}, \ldots, x_{l+u}$ are the unlabeled input points. More precisely, there are given $y_1, \ldots, y_l$ output points, such that $y_j \in \mathcal{Y}_{x_j}$ for all $j = 1, \ldots, l$. Then, for the general data let

$$\mathrm{x} := (x_j)_{j=1}^{l+u}, \quad \mathrm{y} := (y_j)_{j=1}^{l}, \quad \mathrm{z} := \left((x_j)_{j=l+1}^{l+u}, (y_j)_{j=1}^{l}\right).$$

Let W^{l+u} denote the Hilbert space

$$\mathrm{W}^{l+u} = \bigoplus_{j=1}^{l+u} \mathcal{W}_{x_j}. \tag{2.1}$$

For $f \in \mathcal{H}_K$ let

$$\mathrm{f} := (f(x_1), \ldots, f(x_{l+u})) \in \mathrm{W}^{l+u}. \tag{2.2}$$

Also, there is given a (Hermitian, if $\mathbb{F} = \mathbb{R}$) positive semidefinite operator $M \in \mathcal{B}(\mathrm{W}^{l+u})$ represented as an operator block $(l+u) \times (l+u)$-matrix $M = [M_{j,k}]$, with $M_{j,k} \in \mathcal{B}(\mathcal{W}_{x_k}, \mathcal{W}_{x_j})$ for all $j, k = 1, \ldots, l+u$. Let $\mathrm{V} = \{V_x\}_{x \in X}$ be a bundle of maps, loss functions, where $V_x \colon \mathcal{Y}_x \times \mathcal{Y}_x \to \mathbb{R}$ is a function, for all $x \in X$. Also, $\mathrm{C} = \{C_x\}_{x \in X}$ is a bundle of bounded linear operators, where $C_x \colon \mathcal{W}_x \to \mathcal{Y}_x$ for all $x \in X$. The general minimisation problem is

$$f_{\mathrm{z},\gamma} = \operatorname{argmin}_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} f(x_j)) + \gamma_A \|f\|^2_{\mathcal{H}_K} + \gamma_I \langle \mathrm{f}, M\mathrm{f} \rangle_{\mathrm{W}^{l+u}}, \tag{2.3}$$

where $\gamma = (\gamma_A, \gamma_I)$ and $\gamma_A > 0$ and $\gamma_I \geq 0$ are the regularisation parameters. This is a localised version of the general vector valued reproducing kernel Hilbert space for manifold regularised and coregularised multiview learning.

It is also useful to introduce the map to be minimised

$$\mathcal{I}(f) := \frac{1}{l}\sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} f(x_j)) + \gamma_A \|f\|_{\mathcal{H}_K}^2 + \gamma_I \langle \mathrm{f}, M\mathrm{f}\rangle_{\mathrm{W}^{l+u}}$$

and, since $f(x) = K_x^* f$ for all $f \in \mathcal{H}_K$ and all $x \in X$, it equals

$$= \frac{1}{l}\sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} K_{x_j}^* f) + \gamma_A \|f\|_{\mathcal{H}_K}^2 + \gamma_I \langle \mathrm{f}, M\mathrm{f}\rangle_{\mathrm{W}^{l+u}}, \quad (2.4)$$

Reproducing Kernel Hilbert Spaces

We briefly review some concepts and facts on reproducing kernel Hilbert spaces, following classical texts such as:

N. Aronszajn, La théorie générale des noyaux reproduisants et ses applications, Premiére Partie, *Proc. Cambridge Philos. Soc.* **39**(1944), 133–153.

N. Aronszajn, Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**(1950), 337–404.

L. Schwartz, Sous espace Hilbertiens déspaces vectoriel topologiques et noyaux associés (noyaux reproduisants), *J. Analyse Math.* **13**(1964), 115–256.

S. Saitoh, Y. Sawano, *Theory of Reproducing Kernels and Applications*, Springer, Singapore 2016.

V.I. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, Cambridge 2016.

Throughout this presentation we denote by $\mathbb{F}$ one of the commutative fields $\mathbb{R}$ or $\mathbb{C}$. For a nonempty set X let $\mathbb{F}^X$ denote the set of $\mathbb{F}$-valued functions on X, forming an $\mathbb{F}$-vector space under pointwise addition and scalar multiplication. For each $p \in X$, the *evaluation map* at p is the linear functional

$$\mathrm{ev}_p \colon \mathbb{F}^X \to \mathbb{F};\ f \mapsto f(p).$$

The evaluation maps equip $\mathbb{F}^X$ with the locally convex topology of pointwise convergence, which is the weakest topology on $\mathbb{F}^X$ that renders each evaluation map continuous. Under this topology, a generalized sequence in $\mathbb{F}^X$ converges if and only if it converges pointwise, i.e. its image under each evaluation map converges. Since each evaluation map is linear and hence the vector space operations are continuous, this renders $\mathbb{F}^X$ into a complete Hausdorff locally convex space. With respect to this topology, if $\mathcal{X}$ is a topological space, a map $\phi \colon \mathcal{X} \to \mathbb{F}^X$ is continuous if and only if $\mathrm{ev}_p \circ \phi \colon \mathcal{X} \to \mathbb{F}$ is continuous for all $p \in X$.

We are interested in Hilbert spaces $\mathcal{H} \subseteq \mathbb{F}^X$ with topologies at least as strong as the topology of pointwise convergence of $\mathbb{F}^X$, so that the convergence of a sequence of functions in $\mathcal{H}$ implies that the functions also converge pointwise. When X is a finite set, $\mathbb{F}^X \cong \mathbb{F}^d$, where d is the number of elements of X, can itself be made into a Hilbert space with a canonical inner product $\langle f, g \rangle := \sum_{p \in X} f(p)\overline{g(p)}$, or in general by an inner product induced by a positive semidefinite $d \times d$ matrix. This leads to the concept of reproducing kernel Hilbert spaces.
Recalling the F. Riesz's Theorem of representations of bounded linear functionals on Hilbert spaces, if each $\mathrm{ev}_p : \mathcal{H} \to \mathbb{F}$ restricted to $\mathcal{H} \subseteq \mathbb{F}^X$ is continuous, for each $p \in X$, then there exists a unique vector $K_p \in \mathcal{H}$ such that $\mathrm{ev}_p = \langle \cdot, K_p \rangle$. But, since each vector in $\mathcal{H}$ is itself a function $X \to \mathbb{F}$, these vectors altogether define a map $K\colon X \times X \to \mathbb{F}$, $K(p, q) := K_q(p)$. Also, recall that a map $K\colon X \times X \to \mathbb{F}$ is usually called a kernel.

Definition

Let $\mathcal{H} \subseteq \mathbb{F}^X$ be a Hilbert space, $K\colon X \times X \to \mathbb{F}$ a kernel. For each $p \in X$ define $K_p := K(\cdot, p) \in \mathbb{F}^X$. K is said to be a *reproducing kernel* for $\mathcal{H}$, and $\mathcal{H}$ is then said to be a *reproducing kernel Hilbert space* (RKHS), if, for each $p \in X$, we have

(i) $K_p \in \mathcal{H}$;

(ii) $\mathrm{ev}_p = \langle \cdot, K_p \rangle$, that is, for every $f \in \mathcal{H}$ we have $f(p) = \langle f, K_p \rangle$.

The second property is referred to as the *reproducing property* of the kernel K.

We may then summarize the last few paragraphs with the following characterization.

Theorem

Let $\mathcal{H} \subseteq \mathbb{F}^X$ be a Hilbert space. The following assertions are equivalent:

(i) *The canonical injection $i_{\mathcal{H}}\colon \mathcal{H} \to \mathbb{F}^X$ is continuous.*

(ii) *For each $p \in X$, the map $\mathrm{ev}_p\colon \mathcal{H} \to \mathbb{F}$ is continuous.*

(iii) *$\mathcal{H}$ admits a reproducing kernel.*

In that case, the reproducing kernel admitted by the Hilbert space is unique, by the uniqueness of the Riesz representatives K_p of the evaluation maps. We may further apply the reproducing property to each K_q to obtain that $K(p,q) = \langle K_q, K_p \rangle$ for each $p, q \in X$, yielding the following properties:

(i) For each $p \in X$, $K(p,p) = \|K_p\|^2 \geq 0$.

(ii) For each $p, q \in X$, $K(q,p) = \overline{K(p,q)}$ and

$$|K(p,q)|^2 \leq K(p,p)K(q,q). \tag{3.1}$$

(iii) For each $n \in \mathbb{N}$, $(c_i)_{i=1}^n \in \mathbb{F}^n$, $(p_i)_{i=1}^n \in X^n$,

$$\sum_{i=1}^n \sum_{j=1}^n \overline{c_i} c_j K(p_i, p_j) = \| \sum_{i=1}^n c_i K_{p_i} \|^2 \geq 0.$$

The property in (3.1) is the analogue of the Schwarz Inequality. As a consequence of it, if $K(p,p) = 0$ for some $p \in X$ then $K(p,q) = K(q,p) = 0$ for all $q \in X$.

For any $K\colon X \times X \to \mathbb{F}$, each $K_p \in \mathbb{F}^X$ so we may define the subspace

$$\widetilde{\mathcal{H}}_K := \operatorname{span}\{K_p \mid p \in X\}$$

of $\mathbb{F}^X$. If K is the reproducing kernel of a Hilbert space $\mathcal{H}$, $\widetilde{\mathcal{H}}_K$ is also a subspace of $\mathcal{H}$ and

$$\widetilde{\mathcal{H}}_K^{\perp} = \{f \in \mathcal{H} \mid \forall p \in X, f(p) = \langle f, K_p \rangle = 0\} = \{0\},$$

therefore, $\widetilde{\mathcal{H}}_K$ is a dense subspace of $\mathcal{H}$, equivalently, $\{K_p \mid p \in X\}$ is a total set for $\mathcal{H}$.

The property at item (iii) is known as the *positive semidefiniteness property*. A positive semidefinite kernel K is called *definite* if $K(p,p) \neq 0$ for all $p \in X$. Positive semidefiniteness is in fact sufficient to characterize all reproducing kernels.

Theorem (Moore-Aronszajn)

Let $K\colon X \times X \to \mathbb{F}$ be a positive semidefinite kernel. Then there is a unique Hilbert space $\mathcal{H}_K \subseteq \mathbb{F}^X$ with reproducing kernel K.

Let us briefly recall the construction of the Hilbert space $\mathcal{H}_K$ in the proof. We first render $\widetilde{\mathcal{H}}_K$ into a pre-Hilbert space satisfying the reproducing property. Define on $\widetilde{\mathcal{H}}_K$ the inner product

$$\langle \sum_{i=1}^{n} a_i K_{p_i}, \sum_{j=1}^{m} b_j K_{q_j} \rangle_{\widetilde{\mathcal{H}}_K} := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \overline{b_j} K(q_j, p_i)$$

for any $\sum_{i=1}^{n} a_i K_{p_i}, \sum_{j=1}^{m} b_j K_{q_j} \in \widetilde{\mathcal{H}}_K$. It is proven that the definition is correct and provides indeed an inner product.
Let $\widehat{\mathcal{H}}_K$ be the completion of $\widetilde{\mathcal{H}}_K$, then $\widehat{\mathcal{H}}_K$ is a Hilbert space with an isometric embedding $\phi\colon \widetilde{\mathcal{H}}_K \to \widehat{\mathcal{H}}_K$ whose image is dense in $\widehat{\mathcal{H}}_K$. It is proven that this abstract completion can actually be realized in $\mathbb{F}^X$ and that it is the RKHS with reproducing kernel K that we denote by $\mathcal{H}_K$.

In applications, one of the most useful tool is the interplay between reproducing kernels and orthonormal bases of the underlying RKHSs. Although this fact holds in higher generality, we state it for separable Hilbert spaces since, most of the time, this is the case of interest.

Theorem

Let $\mathcal{H} \subseteq \mathbb{F}^X$ be a separable RKHS, with reproducing kernel K, and let $\{\phi_n\}_n$ be an orthonormal basis of $\mathcal{H}$. Then

$$K(p, q) = \sum_{n=1}^{\infty} \phi_n(p)\overline{\phi_n(q)}, \quad p, q \in X,$$

where the series converges absolutely pointwise.

We now recall a useful result on the construction of new RKHSs and positive semidefinite kernels from existing ones. It also shows that the concept of reproducing kernel Hilbert space is actually a special case of the concept of operator range.

Theorem

Let $\mathcal{H}$ be a Hilbert space, $\Phi\colon \mathcal{H} \to \mathbb{F}^X$ a continuous linear operator. Then $\Phi(\mathcal{H}) \subseteq \mathbb{F}^X$ with the norm

$$\|f\|_{\Phi(\mathcal{H})} := \min\{\|u\|_{\mathcal{H}} \mid u \in \mathcal{H},\ f = \Phi u\}$$

is a RKHS, unitarily isomorphic to $(\ker \Phi)^{\perp}$. The kernel for $\Phi(\mathcal{H})$ is then given by the map $(p, q) \mapsto \langle u_q, u_p\rangle = (\mathrm{ev}_p \circ \Phi)(u_q) = \Phi(u_q)(p)$ where $u_q \in \mathcal{H}$ such that $\mathrm{ev}_q \circ \Phi = \langle \cdot, u_q\rangle$ on $\mathcal{H}$.

Applying this proposition to particular continuous linear maps, one obtains useful results for pullbacks, restrictions, sums, scaling, and normalizations of kernels.

We now consider domains equipped with an additional topological or differential structure and recall the relations between the properties of the kernel with respect to this structure to properties of the functions in the corresponding reproducing kernel Hilbert space, e.g. see [15, Section 2.1.3]. Let $\mathcal{H} \subseteq \mathbb{F}^X$ be a RKHS, K its corresponding kernel. Define

$$\Phi_K \colon X \to \mathcal{H};\ p \mapsto K_p. \tag{3.2}$$

Theorem (Boundedness of Kernel)

Let $\mathcal{H} \subseteq \mathbb{F}^X$ be a RKHS, K its corresponding kernel. Then K is bounded iff Φ_K is bounded. In that case every function in $\mathcal{H}$ is bounded, and convergence in $\mathcal{H}$ implies uniform convergence.

Theorem (Continuity of Kernel)

Suppose X is a metric space, let $\mathcal{H} \subseteq \mathbb{F}^X$ be a RKHS, K its corresponding kernel. Then K is (uniformly) continuous iff Φ_K is (uniformly) continuous. In that case every function in $\mathcal{H}$ is (uniformly) continuous, and convergence in $\mathcal{H}$ implies uniform convergence on compact sets.

Theorem (Differentiability of Kernel)

Suppose $X \subseteq \mathbb{R}^d$ is open, let $\mathcal{H} \subseteq \mathbb{F}^X$ be a RKHS, K its corresponding kernel. Then for $j = 1, \ldots, d$, K is continuously differentiable in the j^{th} component of both entries on X if and only if Φ_K is continuously differentiable in the j^{th} component, i.e. the limit

$$\partial_{q_j} K_q := \lim_{h \to 0} \frac{K_{q+he_j} - K_q}{h}$$

exists and is continuous with respect to $q \in X$, where $(e_j)_{j=1}^d$ is the canonical basis for $\mathbb{R}^d$. In that case, every function in $\mathcal{H}$ is once continuously differentiable in the j^{th} component, and we have $(\partial_j f)(q) = \langle f, \partial_{q_j} K_q \rangle$ for each $f \in \mathcal{H}, q \in X$.
The j^{th} partial derivatives of functions in $\mathcal{H}$ are contained in another reproducing kernel Hilbert space $\partial_j \mathcal{H}$, with kernel $\partial_{p_j} \partial_{q_j} K$, such that the map $\partial_j \colon \mathcal{H} \to \partial_j H$ is not only continuous but non-expansive, and unitary if $\mathcal{H}$ does not contain any nonzero function constant in the j^{th} component.

The previous theorem has natural generalizations for functions of class $\mathcal{C}^k(X)$ for $k \geq 1$, and functions that are real or complex analytic on X.

Example (A non-example)

The Banach space $C[0,1]$ has the property that for any $x \in [0,1]$ the evaluation functional $\mathrm{ev}_x \colon C[0,1] \to \mathbb{C}$ is continuous and it is dense in $L^2[0,1]$. However, for any $x \in [0,1]$ the evaluation map ev_x is not continuous with respect to the norm of $L^2[0,1]$.
Indeed, if $0 < x < 1$ then letting

$$f_n(t) := \begin{cases} \frac{t^n}{x^n}, & 0 \leq t \leq x, \\ \frac{(1-t)^n}{(1-x)^n}, & x < t \leq 1, \end{cases}$$

is a sequence of functions in $C[0,1]$ such that $\mathrm{ev}_x(f_n) = f_n(x) = 1$ for all $n \geq 1$ but

$$\lim_{n\to\infty} \|f_n\|_{L^2[0,1]} = 0.$$

For $x = 0$ we can take $f_n(t) := (1-t)^n$ and, for $x = 1$ we can take $f_n(t) := t^n$.
This shows that $L^2[0,1]$ cannot have a structure of a RKHS on $[0,1]$.
In general, RKHS are quite different from L^2-spaces.

Example (Uniform distribution on a compact interval)

Let $(\mu_j)_{j\in\mathbb{Z}} \in l_1(\mathbb{Z})$ be such that $\mu_j > 0$ for all $j \in \mathbb{Z}$ and denote $\mu := \sum_{j\in\mathbb{Z}} \mu_j$. For each $j \in \mathbb{Z}$ define

$$\phi_j \colon [-\pi, \pi] \to \mathbb{C}, \quad \phi_j(t) := \mathrm{e}^{\mathrm{i}\pi j t}, \quad t \in [-\pi, \pi],$$

and consider the Hilbert space

$$\mathcal{H} = \left\{ \sum_{j\in\mathbb{Z}} c_j \phi_j \mid \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j} < \infty \right\},$$

with the inner product

$$\langle \sum_{j\in\mathbb{Z}} c_j \phi_j, \sum_{j\in\mathbb{Z}} d_j \phi_j \rangle = \sum_{j\in\mathbb{Z}} \frac{c_j \overline{d_j}}{\mu_j}.$$

Example

Then $\{\sqrt{\mu_j}\phi_n\}_{j\in\mathbb{Z}}$ is an orthonormal basis of $\mathcal{H}$ and, for an arbitrary function $f \in \mathcal{H}$, we have the Fourier representation

$$f(t) = \sum_{j\in\mathbb{Z}} c_j\phi_j(t), \quad t \in [-\pi, \pi], \tag{4.1}$$

with coefficients $\{c_j\}_{j\in\mathbb{Z}}$ subject to the condition

$$\|f\|_{\mathcal{H}}^2 := \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j} < \infty, \tag{4.2}$$

where the convergence of the series from (8.1) is at least guaranteed with respect to the norm $\|\cdot\|_{\mathcal{H}}$.

Example

However, for any $m \in \mathbb{N}_0$ and $t \in [-\pi, \pi]$, by the Cauchy inequality we have

$$\sum_{|j| \geq m} |c_j \phi_j(t)| \leq \Big(\sum_{|j| \geq m} \frac{|c_j|^2}{\mu_j} \Big)^{1/2} \Big(\sum_{|j| \geq m} \mu_j \Big)^{1/2} \xrightarrow[m \to \infty]{} 0,$$

hence the convergence in (8.1) is absolutely and uniformly on $[-\pi, \pi]$, in particular f is continuous.

Example

By Theorem 16 $\mathcal{H}$ has the reproducing kernel

$$K(s,t) = \sum_{j\in\mathbb{Z}} \mu_j \mathrm{e}^{\mathrm{i}\pi j(s-t)} = \sum_{j\in\mathbb{Z}} \mu_j \phi_j(s)\overline{\phi_j(t)}, \tag{4.3}$$

and the convergence of the series is guaranteed at least pointwise. In addition, for any $t \in [-\pi, \pi]$ we have

$$K(t,t) = \sum_{j\in\mathbb{Z}} \mu_j |\phi_j(t)|^2 = \sum_{j\in\mathbb{Z}} \mu_j = \mu,$$

and hence the kernel K is bounded. In particular, this implies that, actually, the series in (8.3) converges absolutely and uniformly on $[-\pi, \pi]$, hence the kernel K is continuous on $[-\pi, \pi] \times [-\pi, \pi]$. That is, $K(s,t)$ is given by $\kappa(s-t)$ where $\kappa \colon \mathbb{R} \to \mathbb{C}$ is a continuous function with period 2π whose Fourier coefficients $(\mu_j)_{j\in\mathbb{Z}}$ are all positive and absolutely summable.

Example (The Hardy space $H^2(\mathbb{D})$)

We consider the open unit disc in the complex plane $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ and the Szegö kernel

$$K(z,\zeta) = \frac{1}{1 - z\overline{\zeta}} = \sum_{n=0}^{\infty} z^n \overline{\zeta}^n, \quad z, \zeta \in \mathbb{D}, \tag{4.4}$$

where the series converges absolutely and uniformly on any compact subset of $\mathbb{D}$. The RKHS associated to K is the Hardy space $H^2(\mathbb{D})$ of all functions $f : \mathbb{D} \to \mathbb{C}$ that are holomorphic in $\mathbb{D}$ with power series expansion

$$f(z) = \sum_{n=0}^{\infty} f_n z^n, \tag{4.5}$$

such that the coefficients sequence $(f_n)_n$ is in $\ell^2_{\mathbb{C}}(\mathbb{N}_0)$.

Example

The inner product in $H^2(\mathbb{D})$ is

$$\langle \sum_{n=0}^{\infty} f_n z^n, \sum_{n=0}^{\infty} g_n z^n \rangle = \sum_{n=0}^{\infty} f_n \overline{g_n},$$

with norm

$$\| \sum_{n=0}^{\infty} f_n z^n \|^2 = \sum_{n=0}^{\infty} |f_n|^2.$$

For each $\zeta \in \mathbb{D}$ we have

$$\|K_\zeta\| = \left(\sum_{n=0}^{\infty} |\zeta|^{2n}\right)^{1/2} = \frac{1}{\sqrt{1-|\zeta|^2}},$$

hence the kernel K is unbounded.

Example (The Hardy Space $H^2(\mathbb{D}^n)$)

Given $n \in \mathbb{N}$, we let $j = (j_1, \dots, j_n) \in \mathbb{N}_0^n$ be a multi-index and, for $z = (z_1, \dots, z_n) \in \mathbb{C}^n$ we set $z^j := z_1^{j_1} \cdots z_n^{j_n}$. A power series in n variables is a formal expression

$$f(z) = \sum_{j \in \mathbb{N}_0^n} a_j z^j, \tag{4.6}$$

where $a_j \in \mathbb{C}$ for all $j \in \mathbb{N}_0^n$.

The Hardy space $H^2(\mathbb{D}^n)$ is the set of all power series f as in (4.6), where $z \in \mathbb{D}^n$ and

$$\sum_{j \in \mathbb{N}_0^n} |a_j|^2 < +\infty,$$

hence the power series converges absolutely and uniformly on any compact subset in $\mathbb{D}^n$ and defines an analytic function in $\mathbb{D}^n$.

$H^2(\mathbb{D}^n)$ is a RKHS and its kernel is given by

$$K(\zeta, w) = \sum_{j \in \mathbb{N}_0^n} \overline{w}^j \zeta^j = \prod_{k=1}^{n} \frac{1}{1 - \overline{w}_k \zeta_k}, \quad \zeta, w \in \mathbb{D}^n.$$

Examples

Example (Paley-Wiener spaces and kernels)

Let $M > 0$. The Paley-Wiener space with bandwidth M is defined as

$$\mathcal{H}_M := \left\{ f \in L^2(\mathbb{R}) \mid \hat{f} \in L^2[-M, M] \right\},$$

where $\hat{f}$ is the Fourier transform of f, is a closed subspace of $L^2(\mathbb{R})$, and a RKHS with the kernel

$$\begin{aligned} K_M(x, y) &:= \frac{1}{\sqrt{2\pi}} \int_{-M}^{M} \mathrm{e}^{\mathrm{i}\omega(x-y)} \,\mathrm{d}\omega \\ &= \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{\mathrm{i}M(x-y)} - \mathrm{e}^{-\mathrm{i}M(x-y)}}{\mathrm{i}(x-y)} = \sqrt{\frac{2}{\pi}} M \mathrm{sinc}(M(x-y)) \end{aligned}$$

by the unitarity of the Fourier transform, where $\mathrm{sinc}(x) := \frac{\sin x}{x}$ for $x \in \mathbb{R} \setminus \{0\}$ and $\mathrm{sinc}(0) = 1$.

Example

For $M' > M$, $\mathcal{H}_{M'}$ contains $\mathcal{H}_M$ as a closed subspace. By the properties of the Fourier transform and the fact that $\int_{-M}^{M} |\omega|^2 \,\mathrm{d}\omega < \infty$, derivatives of functions in $\mathcal{H}_M$ are also contained in $\mathcal{H}_M$.

By the well-known *Shannon-Nyquist sampling theorem*, $(K_{M,n\pi/M})_{n\in\mathbb{Z}}$ is an orthogonal basis for $\mathcal{H}_M$. This is usually shown by noting the convergence of the Fourier series on $L^2[-M, M]$. We then have the *Whittaker-Shannon interpolation formula*, actually dating back to Émile Borel:

$$f(t) = \sum_{n\in\mathbb{Z}} f(nT) \frac{K_{\pi/T}(t, nT)}{K_{\pi/T}(nT, nT)} = \sum_{n\in\mathbb{Z}} f(nT) \operatorname{sinc}\left(\pi \frac{t - nT}{T}\right)$$

for $f \in \mathcal{H}_M$, $t \in \mathbb{R}$ and $T \leq \pi/M$. That is, a signal of bandwidth M can be interpolated exactly from samples of frequency at least $2M$.

Example (Sobolev spaces)

Recall that, given an interval I in $\mathbb{R}$, a function $f\colon I \to \mathbb{R}$ (or $\mathbb{C}$) is called *absolutely continuous* if, for any $\epsilon > 0$ there exists $\delta > 0$ such that whenever a finite sequence of pairwise disjoint subintervals (x_k, y_k), $k = 1, \ldots, n$ has

$$\sum_{k=1}^{n} (y_k - x_k) < \delta,$$

then

$$\sum_{k}^{n} |f(y_k) - f(x_k)| < \varepsilon.$$

The collection of all absolutely continuous functions on I is denoted by $\mathrm{AC}(I)$.

Example

The following conditions on a function f on a compact interval $[a, b]$ are equivalent:
(1) f is absolutely continuous;
(2) f has a derivative f' almost everywhere, the derivative is Lebesgue integrable, and

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t, \quad x \in [a, b];$$

(3) there exists a Lebesgue integrable function g on $[a, b]$ such that

$$f(x) = f(a) + \int_a^x g(t)\,\mathrm{d}t, \quad x \in [a, b].$$

If these equivalent conditions are satisfied then necessarily $g = f'$ almost everywhere.
Equivalence between (1) and (3) is known as the fundamental theorem of Lebesgue integral calculus, due to Lebesgue.

Example

• The sum and difference of two absolutely continuous functions are also absolutely continuous. If the two functions are defined on a bounded closed interval, then their product is also absolutely continuous.

• If an absolutely continuous function is defined on a bounded closed interval and is nowhere zero then its reciprocal is absolutely continuous.

• Every absolutely continuous function is uniformly continuous and, therefore, continuous. Every Lipschitz-continuous function is absolutely continuous.

• If $f : [a, b] \to \mathbb{R}$ is absolutely continuous, then it is of bounded variation on $[a, b]$.

Example

• If $f : [a, b] \to \mathbb{R}$ is absolutely continuous, then it can be written as the difference of two monotonic nondecreasing absolutely continuous functions on $[a, b]$.

• If $f : [a, b] \to \mathbb{R}$ is absolutely continuous, then it has the Luzin N property, that is, for any Lebesgue negligeable $L \subseteq [a, b]$ the set $f(L)$ is Lebesgue negligeable.

• $f : I \to \mathbb{R}$ is absolutely continuous if and only if it is continuous, is of bounded variation and has the Luzin N property.

Example

Let $\mathcal{H}$ be the set of all absolutely continuous functions $f\colon [0,1] \to \mathbb{R}$ such that its derivative $f' \in L^2[0,1]$ and $f(0) = f(1) = 0$. Then $\mathcal{H}$ is a vector space of functions on $[0,1]$ that is a Hilbert space with respect to the inner product

$$\langle f, g \rangle := \int_0^1 f'(t) g'(t) \, \mathrm{d}t, \quad f, g \in \mathcal{H},$$

and such that, for every $x \in [0,1]$ the evaluation functional ev_x is bounded with $\|\mathrm{ev}_x\| \leq \sqrt{x}$. This follows, e.g. by observing that, for any $0 \leq x \leq 1$ and any $f \in \mathcal{H}$,

$$f(x) = \int_0^x f'(t) \, \mathrm{d}t = \int_0^1 f'(t) \chi_{[0,x]}(t) \, \mathrm{d}t,$$

hence

$$|f(x)| \leq \left(\int_0^1 f'(t)^2 \, \mathrm{d}t \right)^{1/2} \left(\int_0^1 \chi_{[0,x]}(t) \, \mathrm{d}t \right)^{1/2} = \sqrt{x} \|f\|_{L^2[0,1]}.$$

Example

In order to find the reproducing kernel K of $\mathcal{H}$, for arbitrary $f \in \mathcal{H}$ and $x \in [0,1]$, by integration by parts,

$$\begin{aligned} f(x) = \langle f, K_x \rangle &= \int_0^1 f'(t) K_x'(t)\,\mathrm{d}t \\ &= f(t)K_x'(t)\Big|_0^1 - \int_0^1 f(t)K_x''(t)\,\mathrm{d}t = -\int_0^1 f(t)K_x''(t)\,\mathrm{d}t, \end{aligned}$$

thus, K_x is the Green function of the Laplacian with Dirichlet bounday conditions, that is, the solution of the boundary value problem

$$-K_x''(t) = \delta_x(t), \quad K_x(0) = K_x(1) = 0,$$

where δ_x is the formal Dirac function, which yields

$$K_x(t) = K(t,x) = \begin{cases} (1-x)t, & 0 \le t \le x \le 1, \\ (1-t)x, & 0 \le x \le t \le 1. \end{cases}$$

Example

Then,

$$K_x'(t) = \begin{cases} 1 - x, & t < x, \\ -x, & t > x, \end{cases}$$

hence K_x is differentiable except at x (hence almost everywhere) and is equal to the integral of its derivative, hence absolutely continuous, $K_x(0) = K_x(1) = 0$, and K_x' is square-integrable, hence $K_x \in \mathcal{H}$.

Also, for $f \in \mathcal{H}$, we have

$$\begin{aligned} \langle f, K_x \rangle &= \int_0^1 f'(t) K_x'(t) \, \mathrm{d}t \\ &= \int_0^x f'(t)(1-x) \, \mathrm{d}t + \int_x^1 f'(t) \, (-x) \, \mathrm{d}t \\ &= (f(x) - f(0))(1-x) - x(f(1) - f(x)) = f(x), \end{aligned}$$

hence K is the reproducing kernel of $\mathcal{H}$. Also, $\|\mathrm{ev}_x\|^2 = \|K_x\|^2 = K(x,x) = x(1-x)$, hence $\|\mathrm{ev}_x\| = \sqrt{x(1-x)}$.

Integration of RKHS-Valued Functions

Let $(\mathcal{E}; \|\cdot\|)$ be a (real or complex) Banach space and (X, Σ, μ) a finite measure space. On $\mathcal{E}$ we consider the Borel σ-algebra denoted by $\mathcal{B}(\mathcal{E})$. A map $f : X \to \mathcal{E}$ is called *measurable* if $f^{-1}(S) \in \Sigma$ for all $S \in \mathcal{B}(\mathcal{E})$ and it is called *strongly measurable* if it is measurable and its range $f(X)$ is separable. If $\mathcal{E}$ is a separable Banach space then the concepts coincide. Both sets of measurable functions, respectively strongly measurable functions, are vector spaces.

A map $\phi\colon X \to \mathcal{E}$ is *simple* if it is measurable and its range $\phi(X)$ is finite, equivalently, there exist $b_1, \dots, b_n \in \mathcal{E}$ and $E_1, \dots, E_n \in \Sigma$ such that

$$\phi = \sum_{k=1}^{n} b_k \chi_{E_k}, \tag{5.1}$$

where we denote, as usually, by χ_A the characteristic (or indicator) function of A.

It is proven that, a function $f\colon X \to B$ is strongly measurable if and only if there exists a sequence of simple functions $(\phi_n)_n$ such that $\phi_n \underset{n}{\to} f$ pointwise on X. In addition, in this case, the sequence $(\phi_n)_n$ can be chosen such that $\|\phi_n(x)\| \le \|f(x)\|$ for all $x \in X$.

A function $f\colon X \to \mathcal{E}$ is *Bochner integrable* if it is strongly measurable and the scalar function $X \ni x \mapsto \|f(x)\| \in \mathbb{R}$ is integrable. In this case, the Bochner integral of f is defined as follows. Firstly, for a Bochner integrable function ϕ as in (5.1), it is proven that $\mu(E_k) < \infty$ for all $k = 1, \ldots, n$ and then, its Bochner integral is defined by

$$\int_X \phi(x)\,\mathrm{d}\mu(x) := \sum_{k=1}^{n} b_k \mu(E_k) \in \mathcal{E}.$$

In general, if f is Bochner integrable, then there exists a sequence of simple functions $(\phi_n)_n$ that converges pointwise to f on X and $\|\phi_n(x)\| \leq \|f(x)\|$ for all $x \in X$ and all $n \in \mathbb{N}$. In this case, it can be proven that the sequence $(\int_X \phi_n(x)\,\mathrm{d}\mu(x))_n$ is Cauchy in $\mathcal{E}$, hence it has a limit and we define

$$\int_X f(x)\,\mathrm{d}\mu(x) := \lim_{n\to\infty} \int_X \phi_n(x)\,\mathrm{d}\mu(x).$$

It can be proven that this definition is correct, that is, it does not depend on the sequence $(\phi_n)_n$.

Bochner integrable functions share many properties with scalar-valued integrable functions, but not all. For example, the collection of all Bochner integrable functions make a vector space and, for any Bochner integrable function f we have

$$\left\| \int_X f(x)\,\mathrm{d}\mu(x) \right\| \leq \int_X \|f(x)\|\,\mathrm{d}\mu(x). \tag{5.2}$$

Also, letting $L^1(X;\mu;\mathcal{E})$ denote the collection of all equivalence classes of Bochner integrable functions, identified μ-almost everywhere, this is a Banach space with norm

$$\|f\|_1 := \int_X \|f(x)\|\,\mathrm{d}\mu(x), \quad f \in L^1(X;\mu;\mathcal{E}).$$

In addition, the Dominated Convergence Theorem holds for the Bochner integral as well, e.g. see [3, Theorem E.6].

We will use the following result, which is a special case of a theorem of E. Hille, e.g. see [5, Theorem III.2.6]. In Hille's Theorem, the linear transformation is supposed to be only closed and, consequently, additional assumptions are needed, so we provide a proof for the special case of bounded linear operators for the reader's convenience.

Theorem

Let $\mathcal{E}$ be a Banach space, (X, μ) a measure space, and $f : X \to \mathcal{E}$ a Bochner integrable function. If $L \colon \mathcal{E} \to \mathcal{F}$ is a continuous linear transformation between Banach spaces, then $L \circ f : X \to \mathcal{F}$ is Bochner integrable and

$$\int_X (L \circ f)(x) \, \mathrm{d}\mu(x) = L \int_X f(x) \, \mathrm{d}\mu(x).$$

Proof. Since f is Bochner integrable, there exists a sequence $(\phi_n)_n$ of simple functions that converges pointwise to f on X and $\|\phi_n(x)\| \leq \|f(x)\|$ for all $x \in X$ and all $n \in \mathbb{N}$. Then,

$$\|L\phi_n(x) - Lf(x)\| = \|L(\phi_n(x) - f(x))\| \leq \|L\| \|\phi_n(x) - f(x)\| \underset{n}{\to} 0, \quad x \in X$$

hence the sequence $(L \circ \phi_n)_n$ converges pointwise to $L \circ f$. Also, it is easy to see that $L \circ \phi_n$ is a simple function for all $n \in \mathbb{N}$. These show that $L \circ f$ is strongly measurable. Since $\|Lf(x)\| \leq \|L\| \|f(x)\|$ for all $x \in X$ and f is Bochner integrable, it follows that

$$\int_X \|Lf(x)\| \, \mathrm{d}\mu(x) \leq \|L\| \int_X \|f(x)\| \, \mathrm{d}\mu(x) < \infty,$$

hence $L \circ f$ is Bochner integrable.

On the other hand,

$$\|L\phi_n(x)\| \le \|L\|\|\phi_n(x)\| \le \|L\|\|f(x)\|, \quad x \in X,\ n \in \mathbb{N},$$

hence, by the Dominated Convergence Theorem for the Bochner integral, it follows that

$$\begin{aligned}\int_X Lf(x)\,\mathrm{d}\mu(x) &= \lim_{n\to\infty}\int_X L\phi_n(x)\,\mathrm{d}\mu(x) = \lim_{n\to\infty} L\int_X \phi_n(x)\,\mathrm{d}\mu(x)\\ &= L\lim_n \int_X \phi_n(x)\,\mathrm{d}\mu(x) = L\int_X f(x)\,\mathrm{d}\mu(x). \quad \square\end{aligned}$$

A direct consequence of this fact is a sufficient condition for when a pointwise integral coincides with the Bochner integral, valid not only for RKHSs but also for Banach spaces of functions on which evaluation maps at any point are continuous, e.g. $C(Y)$ for some compact Hausdorff space Y.

Proposition

Let (X, Σ, μ) be a measure space, $\mathcal{B} \subseteq \mathbb{F}^X$ a Banach space of functions on X such that all evaluation maps on $\mathcal{B}$ are continuous. Let $\lambda\colon X \times X \to \mathbb{F}$ be such that for each $q \in X$ we have $\lambda_q := \lambda(\cdot, q) \in \mathcal{B}$.
If, for each $q \in X$, the map $X \ni q \mapsto \lambda_q \in \mathcal{B}$ is Bochner integrable, then the scalar map $X \ni q \mapsto \lambda(p, q) \in \mathbb{F}$ is integrable, for each fixed $p \in X$.
Moreover, in that case, the pointwise integral map $X \ni p \mapsto \int_X \lambda(p, q)\, \mathrm{d}\mu(q)$ lies in $\mathcal{B}$ and coincides with the Bochner integral $\int_X \lambda_q \, \mathrm{d}\mu(q)$.

Proof. Since, for each $q \in X$, the map
$X \ni q \mapsto \phi(q) := \lambda(\cdot, q) \in \mathcal{B}$ is Bochner integrable, and taking into account that, for all $p \in X$, the linear functional ev_p is continuous, by Theorem 38 we have

$$\mathrm{ev}_p \int_X \phi(q)\,\mathrm{d}\mu(q) = \int_X \mathrm{ev}_p \circ \phi(q)\,\mathrm{d}\mu(q).$$

Since $\mathrm{ev}_p \circ \phi(q) = \lambda(p, q)$ for all $p, q \in X$, this means that the scalar map $X \ni q \mapsto \lambda(p, q) \in \mathbb{F}$ is integrable, for each fixed $p \in X$, and

$$\mathrm{ev}_p \int_X \phi(q)\,\mathrm{d}\mu(q) = \int_X \lambda(p, q)\,\mathrm{d}\mu(q), \quad p \in X,$$

hence, the pointwise integral map $X \ni p \mapsto \int_X \lambda(p, q)\,\mathrm{d}\mu(q)$ lies in $\mathcal{B}$ and coincides with the Bochner integral $\int_X \lambda_q\,\mathrm{d}\mu(q)$. □

Convergence of Discrete Sampling in RKHSs

Let $(\mathcal{H}, K)$ be a separable RKHS over a set X. Given $f \in \mathcal{H}$ and fixed $(x_i)_{i=1}^N \in X$, the problem of finding the optimal $(\omega_i^N(f))_{i=1}^N \in \mathbb{F}$ to minimize $\|f - \sum_{i=1}^N \omega_i^N(f) K_{x_i}\|_{\mathcal{H}}$ is straightforward: $\sum_{i=1}^N \omega_i^N(f) K_{x_i}$ is the orthogonal projection of f to $\operatorname{span}\{K_{x_i}\}_{i=1}^N$.

We may assume without loss of generality that $\{K_{x_i}\}_{i=1}^N$ are linearly independent, by removing points as necessary without affecting $\operatorname{span}\{K_{x_i}\}_{i=1}^N$ (or losing any information about f, since $\sum_{i=1}^N c_i K_{x_i} = 0$ implies $\sum_{i=1}^N \overline{c_i} f(x_i) = 0$ by the reproducing property).

Proposition

Let $(x_i)_{i=1}^N \in X$ *such that* $\{K_{x_i}\}_{i=1}^N$ *are linearly independent, consider the finite-dimensional subspace* $\mathcal{H}_x^N := \mathrm{span}\{K_{x_i}\}_{i=1}^N$ *of* $\mathcal{H}$*. Then the orthogonal projection* π_x^N *of* $\mathcal{H}$ *onto* $\mathcal{H}_x^N$ *is given by*

$$\pi_x^N(f) = \sum_{i=1}^N \omega_i^\pi(f) K_{x_i} := \sum_{i=1}^N \sum_{j=1}^N f(x_j)\Gamma_{ji}^N K_{x_i} = \sum_{i=1}^N \sum_{j=1}^N \langle f, K_{x_j}\rangle \Gamma_{ji}^N K_{x_i}$$

for any $f \in \mathcal{H}$*, where* $\Gamma^N \in \mathcal{M}_N(\mathbb{F})$ *is the inverse of the Gram matrix* $G^N := [K(x_j, x_i)]_{i,j=1}^N = \left[\langle K_{x_i}, K_{x_j}\rangle\right]_{i,j=1}^N$ *of* $\{x_1, \ldots, x_N\}$*. More generally, if* $\{K_{x_i}\}_{i=1}^N$ *are not linearly independent, for any subset* $s = (x_{i_j})_{j=1}^K$ *such that* $\{K_{x_{i_j}}\}_{i=1}^N$ *form a basis for* $\mathcal{H}_x^N$*, we have* $\mathcal{H}_x^N = \mathcal{H}_s^K$ *and*

$$\pi_x^N = \pi_s^K = \sum_{j=1}^K \sum_{k=1}^K \langle \cdot, K_{x_{i_k}}\rangle \Gamma_{kj}^K K_{x_{i_j}}.$$

Proof. Let $f \in \mathcal{H}$, then $\pi_x^N(f) \in \mathcal{H}_x^N$ so there exist unique $(\omega_i^\pi(f))_{i=1}^N \in \mathbb{F}$ such that $\pi_x^N(f) = \sum_{i=1}^N \omega_i^\pi(f) K_{x_i}$. $(\omega_i^\pi(f))_{i=1}^N$ can then be solved by noting that $f - \pi_x^N(f) \perp H_x^N$ i.e. $f - \pi_x^N(f) \perp K_{x_j}$ for each $j = 1, \ldots, N$:

$$0 = \langle f - \pi_x^N(f), K_{x_j} \rangle = \langle f, K_{x_j} \rangle - \sum_{i=1}^N \omega_i^\pi(f) \langle K_{x_i}, K_{x_j} \rangle$$

$$= f(x_j) - \sum_{i=1}^N \omega_i^\pi(f) G_{ij}^N$$

thus, since $\{K_{x_i}\}_{i=1}^N$ are linearly independent and G^N is invertible, for each $i = 1, \ldots, N$

$$\omega_i^\pi(f) = \sum_{k=1}^N \omega_k^\pi(f) \delta_{ki} = \sum_{k=1}^N \sum_{j=1}^N \omega_k^\pi(f) G_{kj}^N \Gamma_{ji}^N = \sum_{j=1}^N f(x_j) \Gamma_{ji}^N. \quad \square$$

We now follow Saitoh and Sawano [15, 2.4.4] in showing the strong convergence of π_x^N to the identity map as $N \to \infty$ for appropriately chosen $(x_i)_{i=1}^{\infty}$.
Since $\mathcal{H}$ is separable, there exists a countable subset of $\{K_p\}_{p \in X}$ which is total in $\mathcal{H}$; thus, there exists a countable set $F \subseteq X$ such that $\operatorname{span}\{K_x\}_{x \in F}$ is dense in $\mathcal{H}$.

This motivates the following definition:

Definition

A countable subset $\{x_i\}_{i=1}^{\infty}$ of X is called a *uniqueness set* for $\mathcal{H}$ if $\{K_{x_i}\}_{i=1}^{\infty}$ is a total set in $\mathcal{H}$, i.e. for all $f \in \mathcal{H}$, $f(x_i) = 0 \ \forall i \in \mathbb{N}$ implies $f = 0$.

Theorem (Ultimate realization of RKHSs, [15, Theorem 2.33])

Let $(\mathcal{H}, K)$ be a RKHS on X, $\{x_i\}_{i=1}^{\infty}$ a uniqueness set such that $\{K_{x_i}\}_{i=1}^{\infty}$ is linearly independent, G^N the Gram matrix for $\{x_i\}_{i=1}^{N}$, $\Gamma^N = (G^N)^{-1}$. Then for each $f \in \mathcal{H}$,

$$\lim_{N\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_i)\Gamma_{ij}^{N} K_{x_j} = f$$

under the topology of $\mathcal{H}$, with distance decreasing monotonically. Consequently,

$$\langle f, g\rangle = \lim_{N\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_i)\Gamma_{ij}^{N} \overline{g(x_j)}$$

for $f, g \in \mathcal{H}$, and

$$f(x) = \langle f, K_x\rangle = \lim_{N\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_i)\Gamma_{ij}^{N} K(x, x_j)$$

Proof. Since each π_x^N, being a projection, is a continuous linear operator with operator norm 1, and $\text{span}\{K_{x_i}\}_{i=1}^{\infty}$ is dense in $\mathcal{H}$, showing $\lim_{N\to\infty} \pi_x^N f = f$ for $f \in \text{span}\{K_{x_i}\}_{i=1}^{\infty}$ is sufficient.
But for each $f \in \text{span}\{K_{x_i}\}_{i=1}^{\infty}$, since f is a linear combination of finitely many K_{x_i}s, there exists $N_f \in \mathbb{N}$ such that $f \in \text{span}\{K_{x_i}\}_{i=1}^{N_f}$.
Then for each $N \geq N_f$, $\pi_x^N f = f$, so $\lim_{N\to\infty} \pi_x^N f = f$. □

Corollary

Let $\{x_i\}_{i=1}^{\infty}$ *be a uniqueness set for* $(\mathcal{H}, K)$, $\{y_i\}_{i=1}^{\infty}$ *a sequence in* $\mathbb{F}$. *Suppose* $\{y_i\}_{i=1}^{\infty}$ *satisfies*

$$\sup_{N \in \mathbb{N}} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i \Gamma_{ij}^{N} \overline{y_j} < \infty.$$

Then there exists (unique) $F \in \mathcal{H}$ *such that* $F(x_i) = y_i \ \forall i \in \mathbb{N}$.

Proof. Define for each $N \in \mathbb{N}$

$$f_N := \sum_{i=1}^{N} y_i \Gamma_{ij}^{N} K_{x_j} \in \mathcal{H}.$$

Then

$$\|f_N\|^2 = \langle f_N, f_N \rangle = \sum_{i=1}^{N} y_i \Gamma_{ij}^{N} \overline{y_j} \leq \sup_{N \in \mathbb{N}} \sum_{i=1}^{N} y_i \Gamma_{ij}^{N} \overline{y_j} < \infty$$

so $(f_N)_N$ is a bounded sequence in $\mathcal{H}$. Then by the Banach-Alaoglu theorem, it has a weakly convergent subsequence $(f_{N_k})_k$.

Let $F \in \mathcal{H}$ be the weak limit of $(f_{N_k})_k$, $i \in \mathbb{N}$. Then there exists $k_i \in \mathbb{N}$ such that $i \leq N_k$ for every $k \geq k_i$. In that case,

$$(f_{N_k})(x_i) = \sum_{l=1}^{N} \sum_{j=1}^{N} y_l \Gamma_{lj}^N K(x_i, x_j) = \sum_{l=1}^{N} y_l \sum_{j=1}^{N} \Gamma_{lj}^N G_{ji}^N = \sum_{l=1}^{N} y_l \delta_{li} = y_i.$$

That is, $\langle f_{N_k}, K_{x_i} \rangle = y_i$ for each $k \geq k_i$. Then by the weak convergence of $(f_{N_k})_k$ to F,

$$F(x_i) = \langle F, K_{x_i} \rangle = \lim_{k \to \infty} \langle f_{N_k}, K_{x_i} \rangle = y_i. \quad \square$$

Main Results

Throughout this section we consider a probability measure space $(X; \Sigma; P)$ and a RKHS $(\mathcal{H}; \langle\cdot,\cdot\rangle)$ in $\mathbb{F}^X$, with norm denoted by $\|\cdot\|$, such that its reproducing kernel K is measurable. In addition, throughout this section, the reproducing kernel Hilbert space $\mathcal{H}$ is supposed to be separable.

Definition

On the measurable space $(X; \Sigma)$ we define the measure P_K by

$$\mathrm{d}P_K(x) = K(x,x)\,\mathrm{d}P(x), \quad x \in X;$$

more precisely, P_K is the absolutely continuous measure with respect to P such that the function $X \ni x \mapsto K(x,x)$ is the Radon-Nikodym derivative of P_K with respect to P.

With respect to the measure space $(X; \Sigma; P_K)$ we consider the Hilbert space $L^2(X; P_K)$ and first obtain a natural bounded linear operator mapping $L^2(X; P_K)$ to $\mathcal{H}$.

Proposition

With notation and assumptions as before, let $\lambda\colon X \to \mathbb{F}$ be a measurable function such that the integral $\int_X |\lambda(x)|^2 \,\mathrm{d}P_K(x)$ is finite. Then the Bochner integral

$$\int_X \lambda(x) K_x \,\mathrm{d}P(x)$$

exists in $\mathcal{H}$.
In addition, the mapping

$$L^2(X; P_K) \ni \lambda \mapsto L_{P,K}\lambda := \int_X \lambda(x) K_x \,\mathrm{d}P(x) \in \mathcal{H}, \tag{7.1}$$

is a nonexpansive, hence bounded, linear operator.

Proof. By assumptions, the map $X \ni x \mapsto \lambda(x)K_x \in \mathcal{H}$ is measurable and, since $\mathcal{H}$ is separable, it follows that this map is actually strongly measurable. Letting $\|\cdot\|$ denote the norm on $\mathcal{H}$ and using the assumption that $\int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}P(x)$ is finite, we have

$$\int_X \|\lambda(x)K_x\|^2\,\mathrm{d}P(x) = \int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}P(x) < \infty,$$

hence, by the Schwarz Inequality and taking into account that P is a probability measure, we have

$$\int_X \|\lambda(x)K_x\|\,\mathrm{d}P(x) \leq \sqrt{\int_X \|\lambda(x)K_x\|^2\,\mathrm{d}P(x)} < \infty.$$

By Theorem 38 this implies that the Bochner integral $\int_X \lambda(x)K_x\,\mathrm{d}P(x)$ exists in $\mathcal{H}$. Consequently, the mapping $L_{P,K}$ as in (7.1) is correctly defined and it is clear that it is a linear transformation.

For arbitrary $\lambda \in L^2(X; P_K)$, by the triangle inequality for the Bochner integral (5.2) we then have

$$\begin{aligned}\left\| \int_X \lambda(x) K_x \, \mathrm{d}P(x) \right\|^2 &\leq \left(\int_X \|\lambda(x) K_x\| \, \mathrm{d}P(x) \right)^2 \\ &= \left(\int_X |\lambda(x)| K(x,x)^{1/2} \, \mathrm{d}P(x) \right)^2\end{aligned}$$

and applying the Schwarz Inequality for the integral and taking into account that P is a probability measure

$$\leq \int_X |\lambda(x)|^2 K(x,x) \, \mathrm{d}P(x) = \|\lambda\|^2_{L^2(X;P_K)},$$

hence $L_{P,K} \colon L^2(X; P_K) \to \mathcal{H}$ is a nonexpansive linear operator. □

Using the bounded linear operator $L_{P,K}$ defined as in (7.1), let us denote its range by

$$\mathcal{H}_P := L_{P,K}(L^2(X; P_K)), \tag{7.2}$$

which is a subspace of the RKHS $\mathcal{H}$.

Proposition

$\mathcal{H}_P$ is a RKHS contained in $\mathcal{H}$, hence in $\mathbb{F}^X$, and its reproducing kernel K_P is

$$K_P(x, y) = \int_X \frac{K(x, z)K(z, y)}{K(z, z)} \, \mathrm{d}P(z), \quad x, y \in X,$$

where, whenever $K(z, z) = 0$, by convention we define $K(x, z)K(z, y)/K(z, z) = 0$ for all $x, y \in X$.

Proof. Since $L^2(X; P_K)$ is a Hilbert space and $L_{P,K}$ is a bounded linear map, by Theorem 17 it follows that $\mathcal{H}_P$ is a RKHS in $\mathbb{F}^X$, isometrically isomorphic to the orthogonal complement of $\ker L_{P,K} \subseteq L^2(X; P_K)$, and its norm is given by

$$\|g\|_{\mathcal{H}_P} := \min\left\{\|\lambda\|_{L^2(X;P_K)} \mid L_{P,K}\lambda = g\right\}, \quad g \in \mathcal{H}_P.$$

Let

$$X_0 := \{x \in X \mid K(x,x) = 0\},$$

and let us define $u_x \colon X \to \mathbb{F}$ by

$$u_x(y) := \begin{cases} \frac{K(y,x)}{K(y,y)}, & y \in X \setminus X_0, \\ 0, & y \in X_0. \end{cases}$$

From the Schwarz Inequality for the kernel K, it follows that if $x \in X_0$ then $K(x,y) = 0$ for all $y \in X$. This shows that $u_x = 0$ for all $x \in X_0$.

For each $x \in X$, by the Schwarz inequality and the fact that P is a probability measure we have

$$\begin{aligned}\int_X |u_x(y)|^2 K(y,y)\,\mathrm{d}P(y) &= \int_{X\setminus X_0} \frac{|K(y,x)|^2}{K(y,y)}\,\mathrm{d}P(y)\\ &\leq \int_{X\setminus X_0} \frac{K(y,y)K(x,x)}{K(y,y)}\,\mathrm{d}P(y)\\ &= K(x,x)P(X\setminus X_0) < \infty,\end{aligned}$$

hence, $u_x \in L^2(X, P_K)$. Then, taking into account that $K(x,y) = 0$ for all $y \in X_0$ and all $x \in X$, it follows that, for each $\lambda \in L^2(X, P_K)$ and $x \in X$, we have

$$\begin{aligned}(L_{P,K}\lambda)(x) &= \int_X \lambda(y)K(x,y)\,\mathrm{d}P(y) = \int_{X\setminus X_0} \lambda(y)K(x,y)\,\mathrm{d}P(y)\\ &= \int_{X\setminus X_0} \lambda(y)\overline{\frac{K(y,x)}{K(y,y)}}K(y,y)\,\mathrm{d}P(y)\\ &= \int_X \lambda(y)\overline{u_x(y)}K(y,y)\,\mathrm{d}P(y) = \langle \lambda, u_x\rangle_{L^2(X,P_K)}.\end{aligned}$$

In conclusion, u_x is exactly the representative for the functional $\mathrm{ev}_x L_{P,K}$ so, by Theorem 17 the kernel of $\mathcal{H}_P$ is

$$\begin{aligned} K_P(x,y) &= \langle u_y, u_x\rangle_{L^2(X,P_K)} \\ &= \int_X u_y(z)\overline{u_x(z)}K(z,z)\,\mathrm{d}P(z) \\ &= \int_{X\setminus X_0} u_y(z)\overline{u_x(z)}K(z,z)\,\mathrm{d}P(z) \end{aligned}$$

and, using the convention that $K(x,z)K(z,y)/K(z,z) = 0$ whenever $K(z,z) = 0$ and for arbitrary $x, y \in X$,

$$= \int_X \frac{K(x,z)K(z,y)}{K(z,z)}\,\mathrm{d}P(z). \quad \square$$

The first step in our enterprise is to find error bounds for approximations of functions in the reproducing kernel Hilbert space $\mathcal{H}$ in terms of distributional finite linear combinations of functions of type K_x.

Theorem

With notation and assumptions as before, let $\lambda \in L^2(X; P_K)$ and $f \in \mathcal{H}$. For each $n \in \mathbb{N}$ and $\delta > 0$, consider the set

$$A_{n,\delta} := \left\{(x_1, \dots, x_n) \in X^n \mid \left\| f - \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i) K_{x_i} \right\|_{\mathcal{H}} \geq \delta \right\}. \quad (7.3)$$

Then, letting P^n denote the product probability measure on X^n and defining the bounded linear operator $L_{P,K}$ as in (7.1), we have

$$P^n(A_{n,\delta}) \leq \frac{1}{\delta^2} \|f - L_{P,K}\lambda\|_{\mathcal{H}}^2 + \frac{1}{n\delta^2} \left(\|\lambda\|_{L^2(X;P_K)}^2 - \|L_{P,K}\lambda\|_{\mathcal{H}}^2 \right).$$

Compare with Theorem 8 and Theorem 11 of Kon-Raphael.

Proof. By Proposition 45, the Bochner integral $\int_X \lambda(x) K_x \, \mathrm{d}P(x)$ exists in $\mathcal{H}$ and the linear operator $L_{P,K}$ is well-defined and bounded. In order to simplify the notation, considering $g\colon X^n \to \mathcal{H}$ the function defined by

$$g(x_1, \ldots, x_n) = f - \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i) K_{x_i}, \quad (x_1, \ldots, x_n) \in X^n,$$

observe that g is measurable and for each $\delta > 0$ we have

$$A_{n,\delta} = \{(x_1, \ldots, x_n) \in X^n \mid \|g(x_1, \ldots, x_n)\| \geq \delta\}. \tag{7.4}$$

Then we have

$$\begin{aligned}\|g(x_1, \ldots, x_n)\|^2 &= \Big\|f - \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i) K_{x_i}\Big\|^2 \\ &= \|f\|^2 - \frac{2}{n} \sum_{i=1}^{n} \mathrm{Re}\langle f, \lambda(x_i) K_{x_i}\rangle \qquad (7.5) \\ &\quad + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \lambda(x_i) K_{x_i}, \lambda(x_j) K_{x_j}\rangle.\end{aligned}$$

Having in mind the Markov-Bienaymé-Chebyshev Inequality as in (1.1), we have to integrate $\|g\|^2$ with respect to the probability product measure P^n.

Since P^n is a probability measure we have

$$\int_{X^n} \|f\|^2 \, \mathrm{d}P^n(x_1, \ldots, x_n) = \|f\|^2.$$

On the other hand, by Fubini's theorem and the fact that the Bochner integral commutes with continuous linear operations, see Theorem 38, we have

$$\begin{aligned}
\int_{X^n} \mathrm{Re}\langle f, \lambda(x_i) K_{x_i} \rangle \, \mathrm{d}P^n(x_1, \ldots, x_n) &= \mathrm{Re}\langle f, \int_{X^n} \lambda(x_i) K_{x_i} \, \mathrm{d}P^n(x_1, \ldots, x_n) \\
&= \mathrm{Re}\langle f, \int_X \lambda(x) K_x \, \mathrm{d}P(x) \rangle \\
&= \mathrm{Re}\langle f, L_{P,K} \lambda \rangle.
\end{aligned}$$

Also, for each $i = 1, \ldots, n$,

$$\int_{X^n} \langle \lambda(x_i) K_{x_i}, \lambda(x_i) K_{x_i} \rangle \, \mathrm{d}P^n(x_1, \ldots, x_n)$$

$$= \int_{X^n} |\lambda(x_i)|^2 K(x_i, x_i) \, \mathrm{d}P^n(x_1, \ldots, x_n)$$

$$= \int_X |\lambda(x)|^2 K(x, x) \, \mathrm{d}P(x),$$

and, for each $i, j = 1, \ldots, n$, $i \neq j$,

$$\int_{X^n} \langle \lambda(x_i) K_{x_i}, \lambda(x_j) K_{x_j} \rangle \, \mathrm{d}P^n(x_1, \ldots, x_n)$$

$$= \int_X \langle \lambda(x_i) K_{x_i}, \int_X \lambda(x_j) K_{x_j} \, \mathrm{d}P(x_j) \rangle \, \mathrm{d}P(x_i)$$

$$= \langle \int_X \lambda(x) K_x \, \mathrm{d}P(x), \int_X \lambda(x) K_x \, \mathrm{d}P(x) \rangle$$

$$= \| \int_X \lambda(x) K_x \, \mathrm{d}P(x) \|^2.$$

Integrating both sides of (7.5) and using all the previous equalities, we therefore have

$$\int_{X^n} \|g(x_1,\dots,x_n)\|^2 \,\mathrm{d}P^n(x_1,\dots,x_n) = \|f\|^2 - \frac{2}{n}\sum_{i=1}^{n} \mathrm{Re}\langle f, \int_X \lambda(x) K_x \,\mathrm{d}P$$

$$+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j=1}^{n} \| \int_X \lambda(x) K_x \,\mathrm{d}P(x)\|^2 + \frac{1}{n^2}\sum_{i=1}^{n}\int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}$$

$$= \|f\|^2 - 2\mathrm{Re}\langle f, \int_X \lambda(x) K_x \,\mathrm{d}P(x)\rangle + \frac{n-1}{n}\| \int_X \lambda(x) K_x \,\mathrm{d}P(x)\|^2$$

$$+ \frac{1}{n}\int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}P(x)$$

$$= \left\| f - \int_X \lambda(x) K_x \,\mathrm{d}P(x)\right\|^2 + \frac{1}{n}\Big(\int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}P(x) - \| \int_X \lambda(x)$$

$$= \|f - L_{P,K}\lambda\|^2 + \frac{1}{n}\Big(\int_X |\lambda(x)|^2 K(x,x)\,\mathrm{d}P(x) - \|L_{P,K}\lambda\|^2\Big).$$

Finally, in view of the Markov-Bienaymé-Chebyshev Inequality as in (1.1), when X is replaced by X^n and P by P^n, and taking into account the previous equality and (7.4), we get

$$\begin{aligned} P^n(A_{n,\delta}) &\leq \frac{1}{\delta^2} \int_{X^n} \|g(x_1, \ldots, x_n)\|^2 \,\mathrm{d}P^n(x_1, \ldots, x_n) \\ &= \frac{1}{\delta^2} \|f - L_{P,K}\lambda\|^2 + \frac{1}{n\delta^2} \left(\|\lambda\|^2_{L^2(X;P_K)} - \|L_{P,K}\lambda\|^2 \right), \end{aligned}$$

which is the required inequality. □

As with the special case of kernel embeddings, for which $\lambda = 1$, see Smola et al. [17], we may use the bound in Theorem 47 to obtain a statement of convergence in probability.

Theorem (Convergence in Probability of Projections)

Let X, P, K, and $\mathcal{H}$ be as in Theorem 47. For each sequence $x = (x_i)_i \in X^{\mathbb{N}}$ and each $n \in \mathbb{N}$, let π_x^n denote the orthogonal projection of $\mathcal{H}$ onto $\operatorname{span}\{K_{x_i}\}_{i=1}^n$. Let $f \in \mathcal{H}$ and, for each $\delta > 0$ and $n \in \mathbb{N}$, define

$$B_{n,\delta} := \{(x_1, \ldots, x_n) \in X^n \mid \|f - \pi_x^n f\| \geq \delta\}.$$

Theorem

Then, for each $\delta > 0$

$$\limsup_{n\to\infty} P^n(B_{n,\delta}) \leq \frac{1}{\delta^2} d_{\mathcal{H}}(f, \mathcal{H}_P)^2,$$

where $d_{\mathcal{H}}(f, \mathcal{H}_P) = \inf_{g\in\mathcal{H}_P} \|f - g\|$.
In particular, if f *belongs to* $\overline{\mathcal{H}_P}^{\mathcal{H}}$, *the closure of* $\mathcal{H}_P$ *with respect to the topology of* $\mathcal{H}$, *then*

$$\lim_{n\to\infty} P^n(B_{n,\delta}) = 0.$$

Proof. Let $\lambda \in L^2(X, P_K)$ and fix $\delta > 0$, arbitrary. Then

$$\|f - \pi_x^n f\| \le \left\| f - \frac{1}{n}\sum_{i=1}^{n} \lambda(x_i) K_{x_i} \right\|, \tag{7.6}$$

hence, with notation as in (7.3), we have $B_{n,\delta} \subseteq A_{n,\delta}$. By Theorem 47, this implies

$$P^n(B_{n,\delta}) \le \frac{1}{\delta^2}\|f - L_{P,K}\lambda\|^2 + \frac{1}{n\delta^2}\left[\|\lambda\|^2_{L^2(X,P_K)} - \|L_{P,K}\lambda\|^2\right].$$

Therefore,

$$\begin{aligned}\limsup_{n\to\infty} P^n(B_{n,\delta}) &\le \limsup_{n\to\infty}\left[\frac{1}{\delta^2}\|f - L_{P,K}\lambda\|^2_{\mathcal{H}} + \frac{1}{n\delta^2}\left(\|\lambda\|^2_{L^2(X,P_K)} - \|L_{P,\dots}\right.\right. \\ &= \frac{1}{\delta^2}\|f - L_{P,K}\lambda\|^2_{\mathcal{H}}.\end{aligned}$$

Thus, since the left-hand side is independent of λ,

$$\limsup_{n\to\infty} P^n(B_{n,\delta}) \le \inf_{\lambda\in L^2(X,P_K)} \frac{1}{\delta^2}\|f - L_{P,K}\lambda\|^2_{\mathcal{H}} = \frac{1}{\delta^2} d_{\mathcal{H}}(f, \mathcal{H}_P)^2.$$

In particular, if f belongs to $\overline{\mathcal{H}_P}^{\mathcal{H}}$, then $d_{\mathcal{H}}(f, \mathcal{H}_P) = 0$. $\square$

In fact, by noting that $\|f - \pi_x^n f\|$, unlike $\left\|f - \frac{1}{n}\sum_{i=1}^n \lambda(x_i)K_{x_i}\right\|$, is monotonically nonincreasing with respect to n by Theorem 42, we can strengthen the preceding statement to almost certain convergence after passing to a single measure space.
Firstly, recall that, e.g. see [3, Proposition 10.6.1], the countably infinite product space $X^{\mathbb{N}}$ equipped with the smallest σ-algebra rendering each projection map $X_i \colon X^{\mathbb{N}} \to X$ measurable admits a unique probability measure $P^{\mathbb{N}}$ such that the projection maps are independent random variables with distribution P.

Lemma

Let X, P, K, and $\mathcal{H}$ be as in Theorem 47 and $f \in \mathcal{H}$. For each $\delta > 0$ define

$$S_{n,\delta} := \left\{ x = (x_k)_{k=1}^{\infty} \in X^{\mathbb{N}} \mid \|f - \pi_x^n f\| \geq \delta \right\}, \quad n \in \mathbb{N},$$

and

$$S_\delta := \left\{ x = (x_k)_{k=1}^{\infty} \in X^{\mathbb{N}} \mid \forall N \in \mathbb{N}, \exists n \geq N, \|f - \pi_x^n f\| \geq \delta \right\} \tag{7.7}$$

$$= \bigcap_{N \in \mathbb{N}} \bigcup_{n \geq N} S_{n,\delta}.$$

Then,

$$P^{\mathbb{N}}(S_\delta) \leq \frac{1}{\delta^2} d_{\mathcal{H}}(f, \mathcal{H}_P)^2,$$

and, consequently, if $f \in \overline{\mathcal{H}_P}^{\mathcal{H}}$, then

$$P^{\mathbb{N}}(S_\delta) = 0.$$

Proof. Observe that for each $n, m \in \mathbb{N}$ such that $n > m$, $\|f - \pi_x^n f\| \leq \|f - \pi_x^m f\|$, for each $x \in X^{\mathbb{N}}$, and hence $S_{n,\delta} \subseteq S_{m,\delta}$ for each $\delta > 0$. Then,

$$S_\delta = \bigcap_{N \in \mathbb{N}} \bigcup_{n \geq N} S_{n,\delta} = \bigcap_{N \in \mathbb{N}} S_{N,\delta},$$

hence, for any $\lambda \in L^2(X, P_K)$,

$$P^{\mathbb{N}}(S_\delta) \leq \inf_{N \in \mathbb{N}} P^{\mathbb{N}}(S_{N,\delta}) \leq \frac{1}{\delta^2} \|f - L_{P,K}\lambda\|_{\mathcal{H}}^2,$$

since $P^{\mathbb{N}}$ is monotone and $S_\delta \subseteq S_{N,\delta}$ for all $N \in \mathbb{N}$. □

Theorem (Almost Certain Convergence of Projections)

Let $X, P, K, \mathcal{H}$ be as in Theorem 47 and suppose $\mathcal{H}_P$ is dense in $\mathcal{H}$. Then, for each $f \in \mathcal{H}$,

$$P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \underset{n}{\to} f\right\}\right) = 1,$$

hence,

$$P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \forall f \in \mathcal{H}, \pi_x^n f \underset{n}{\to} f\right\}\right) = 1.$$

Proof. Let $f \in \mathcal{H}$. With the same sets S_δ defined in (7.7),

$$\left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \not\to f\right\} = \left\{x \in X^{\mathbb{N}} \mid \exists \delta > 0, \forall N \in \mathbb{N}, \exists n \geq N, \|f - \pi_x^n f\| \geq \right.$$

$$= \bigcup_{\delta > 0} S_\delta.$$

Observe further that $S_\delta \subseteq S_{\delta'}$ whenever $\delta > \delta'$, and for each $\delta > 0$ there exists $m \in \mathbb{N}$ such that $\delta > 1/m$, so that

$$\left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \underset{n}{\not\to} f\right\} = \bigcup_{0 < \delta \leq 1} S_\delta = \bigcup_{m \in \mathbb{N}} S_{1/m}$$

thus

$$P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \underset{n}{\not\to} f\right\}\right) \leq \sum_{m \in \mathbb{N}} P^{\mathbb{N}}(S_{1/m}) = \sum_{m \in \mathbb{N}} 0 = 0.$$

Since $\mathcal{H}$ is separable let $\mathcal{D}$ be a countable dense subset of $\mathcal{H}$. Since each π_x^n is a continuous linear operator with operator norm 1, $\pi_x^n f \to f$ for all $f \in \mathcal{H}$ iff $\pi_x^n f \to f$ for all $f \in \mathcal{D}$. Thus by the countable subadditivity of $P^{\mathbb{N}}$,

$$\begin{aligned}
P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \exists f \in \mathcal{H},\ \ \pi_x^n f \underset{n}{\not\to} f\right\}\right) \\
&= P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \exists f \in \mathcal{D}, \pi_x^n f \underset{n}{\not\to} f\right\}\right) \\
&= P^{\mathbb{N}}\left(\bigcup_{f \in \mathcal{D}} \left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \underset{n}{\not\to} f\right\}\right) \\
&\leq \sum_{f \in \mathcal{D}} P^{\mathbb{N}}\left(\left\{x \in X^{\mathbb{N}} \mid \pi_x^n f \underset{n}{\not\to} f\right\}\right) = 0. \quad \square
\end{aligned}$$

In summary, for a given probability measure P under the assumption that it renders the space $\mathcal{H}_P$, the image of $L_{P,K}$, dense in $\mathcal{H}$, a sequence of points sampled independently from P yields a uniqueness set with probability 1.
As a final result, in the next proposition we show a sufficient condition, valid for many applications, when this assumption holds.

Proposition

Let X be a topological space, P a Borel probability measure on X, $\mathcal{H} \subseteq \mathbb{F}^X$ a RKHS with measurable kernel K, and let P_K, $L_{P,K}$ and $\mathcal{H}_P$ defined as in Definition 44, (7.1), and (7.2), respectively. Suppose that K is continuous on X, that $\mathcal{H} \subseteq L^2(X; P_K)$, and that P is strictly positive on any nonempty open subset of X. Then $\mathcal{H}_P$ is dense in $\mathcal{H}$.

Proof. Let $f \in \mathcal{H}$, $f \perp \mathcal{H}_P$. That is, for each $\lambda \in L^2(X; P_K)$, we have

$$\langle f, L_{P,K}\lambda\rangle_{\mathcal{H}} = \langle f, \int_X \lambda(x) K_x \,\mathrm{d}P(x)\rangle = 0.$$

Then noting the fact that $\int \lambda(x) K_x \,\mathrm{d}P(x)$ is a Bochner integral and hence, by Theorem 38, it commutes with inner products,

$$0 = \langle f, \int_X \lambda(x) K_x \,\mathrm{d}P(x)\rangle = \int_X \overline{\lambda(y)}\langle f, K_x\rangle \,\mathrm{d}P(x) = \int \overline{\lambda(x)} f(x) \,\mathrm{d}P(x).$$

By assumption, $f \in \mathcal{H} \subseteq L^2(X; P_K)$, so we can take $\lambda = f$ to obtain

$$\int |f(x)|^2 \,\mathrm{d}P(x) = \int_X \overline{f(x)} f(x) \,\mathrm{d}P(x) = 0.$$

This implies that $f = 0$ P-almost everywhere, i.e. the set $f^{-1}(\mathbb{F} \setminus \{0\})$ has zero P measure.
Since K is continuous by assumption, by Theorem 19 each $f \in \mathcal{H}$ is continuous hence $f^{-1}(\mathbb{F} \setminus \{0\})$ is an open subset of X. But, since P is assumed strictly positive on any nonempty open set, it follows that $f^{-1}(\mathbb{F} \setminus \{0\})$ must be empty, hence $f = 0$ identically. $\square$

Example: Uniform distribution on a compact interval

Let $(\mu_j)_{j\in\mathbb{Z}} \in l_1(\mathbb{Z})$ be such that $\mu_j > 0$ for all $j \in \mathbb{Z}$ and denote $\mu := \sum_{j\in\mathbb{Z}} \mu_j$. For each $j \in \mathbb{Z}$ define

$$\phi_j\colon [-\pi,\pi] \to \mathbb{C}, \quad \phi_j(t) := \mathrm{e}^{\mathrm{i}\pi jt}, \quad t \in [-\pi,\pi],$$

and consider the Hilbert space

$$\mathcal{H} = \left\{ \sum_{j\in\mathbb{Z}} c_j\phi_j \mid \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j} < \infty \right\},$$

with the inner product

$$\langle \sum_{j\in\mathbb{Z}} c_j\phi_j, \sum_{j\in\mathbb{Z}} d_j\phi_j \rangle = \sum_{j\in\mathbb{Z}} \frac{c_j\overline{d_j}}{\mu_j}.$$

Then $\{\sqrt{\mu_j}\phi_j\}_{j\in\mathbb{Z}}$ is an orthonormal basis of $\mathcal{H}$ and, for an arbitrary function $f \in \mathcal{H}$, we have the Fourier representation

$$f(t) = \sum_{j\in\mathbb{Z}} c_j \phi_j(t), \quad t \in [-\pi, \pi], \tag{8.1}$$

with coefficients $\{c_j\}_{j\in\mathbb{Z}}$ subject to the condition

$$\|f\|_{\mathcal{H}}^2 := \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j} < \infty, \tag{8.2}$$

where the convergence of the series from (8.1) is at least guaranteed with respect to the norm $\|\cdot\|_{\mathcal{H}}$. However, for any $m \in \mathbb{N}_0$ and $t \in [-\pi, \pi]$, by the Cauchy inequality we have

$$\sum_{|j|\geq m} |c_j \phi_j(t)| \leq \Big(\sum_{|j|\geq m} \frac{|c_j|^2}{\mu_j}\Big)^{1/2} \Big(\sum_{|j|\geq m} \mu_j\Big)^{1/2} \xrightarrow[m\to\infty]{} 0,$$

hence the convergence in (8.1) is absolutely and uniformly on $[-\pi, \pi]$, in particular f is continuous.

By (**??**) $\mathcal{H}$ has the reproducing kernel

$$K(s,t) = \sum_{j\in\mathbb{Z}} \mu_j \mathrm{e}^{\mathrm{i}\pi j(s-t)} = \sum_{j\in\mathbb{Z}} \mu_j \phi_j(s)\overline{\phi_j(t)}, \qquad (8.3)$$

and the convergence of the series is guaranteed at least pointwise. In addition, for any $t \in [-\pi, \pi]$ we have

$$K(t,t) = \sum_{j\in\mathbb{Z}} \mu_j |\phi_j(t)|^2 = \sum_{j\in\mathbb{Z}} \mu_j = \mu,$$

and hence the kernel K is bounded. In particular, this implies that, actually, the series in (8.3) converges absolutely and uniformly on $[-\pi, \pi]$, hence the kernel K is continuous on $[-\pi, \pi] \times [-\pi, \pi]$. That is, $K(s,t)$ is given by $\kappa(s-t)$ where $\kappa\colon \mathbb{R} \to \mathbb{C}$ is a continuous function with period 2π whose Fourier coefficients $(\mu_j)_{j\in\mathbb{Z}}$ are all positive and absolutely summable.

Let P be the normalized Lebesgue measure on $[-\pi, \pi]$, equivalently, the uniform probability distribution on $[-\pi, \pi]$, and observe that $\{\phi_j\}_{j\in\mathbb{Z}}$ is an orthonormal basis of the Hilbert space $L_P[-\pi, \pi]$. With notation as in (**??**), we have $\mathrm{d}P_K(t) = K(t,t)\,\mathrm{d}P(t) = \mu\,\mathrm{d}P(t)$ hence $L^2_{P_K}[-\pi, \pi] = L^2_P[-\pi, \pi]$ with norms differing by multiplication with $\mu > 0$. In particular, $\{\phi_j/\sqrt{\mu}\}_{j\in\mathbb{Z}}$ is an orthonormal basis of the Hilbert space $L^2_{P_K}[-\pi, \pi]$.

We consider now the nonexpansive operator $L_{P,K} \colon L^2_{P_K}[-\pi, \pi] \to \mathcal{H}$ defined as in (7.1). Then, for any $j \in \mathbb{Z}$ and $t \in [-\pi, \pi]$, we have

$$\begin{aligned}(L_{P,K}\phi_j)(t) &= \int_{-\pi}^{\pi} \phi_j(s) K(t,s) \,\mathrm{d}P(s) = \int_{-\pi}^{\pi} \phi_j(s) \left(\sum_{k \in \mathbb{Z}} \mu_k \phi_k(t) \overline{\phi_k(s)} \right) \\ &= \sum_{k \in \mathbb{Z}} \mu_k \phi_k(t) \int_{-\pi}^{\pi} \phi_j(s) \overline{\phi_k(s)} \,\mathrm{d}P(s) = \sum_{k \in \mathbb{Z}} \mu_k \phi_k(t) \delta_{jk} = \mu_j \phi\end{aligned}$$

where, the series commutes with the integral either by the Bounded Convergence Theorem for the Lebesgue integral, or by using the uniform convergence of the series and the Riemann integral.

Similarly, the Hilbert space $\mathcal{H}_P := L_{P,K}(L^2_{P_K}[-\pi,\pi])$, as in Proposition 46, is a RKHS, with kernel,

$$
\begin{aligned}
K_P(s,t) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\left(\sum_{j\in\mathbb{Z}} \mu_j \phi_j(s)\overline{\phi_j(z)}\right)\left(\sum_{l\in\mathbb{Z}} \mu_l \phi_l(z)\overline{\phi_l(t)}\right)}{\sum_{j\in\mathbb{Z}} \mu_j}\,\mathrm{d}z \\
&= \frac{1}{\mu}\sum_{j\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} \mu_j\mu_l\phi_j(s)\overline{\phi_l(t)}\frac{1}{2\pi}\int_{-\pi}^{\pi}\phi_j(z)\overline{\phi_l(z)}\,\mathrm{d}z \\
&= \frac{1}{\mu}\sum_{j\in\mathbb{Z}}\sum_{l\in\mathbb{Z}} \mu_j\mu_l\phi_j(s)\overline{\phi_l(t)}\delta_{jl} = \sum_{j\in\mathbb{Z}} \frac{\mu_j^2}{\mu}\phi_j(s)\overline{\phi_j(t)}.
\end{aligned}
$$

Thus, letting $\mu_j' := \frac{\mu_j^2}{\mu} \leq \mu_j$, $j\in\mathbb{Z}$ and noting that $\sum_{j\in\mathbb{Z}}\mu_j' \leq \sum_{j\in\mathbb{Z}}\mu_j < \infty$, we have

$$
\mathcal{H}_P = \left\{ \sum_{j\in\mathbb{Z}} c_j\phi_j \mid \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j'} < \infty \right\} = \left\{ \sum_{j\in\mathbb{Z}} c_j\phi_j \mid \sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j^2} < \infty \right\}.
$$

In particular, $\mathcal{H}_P$ is dense in $\mathcal{H}$ since both contain $\mathrm{span}\{\phi_j\}_{j\in\mathbb{Z}}$ as dense subsets, but this follows from the more general statement in Proposition 52 as well.

Let now $\lambda \in L^2_{P_K}[-\pi, \pi] = L^2_P[-\pi, \pi]$ be arbitrary, hence

$$\lambda = \sum_{j \in \mathbb{Z}} \lambda_j \phi_j, \quad \sum_{j \in \mathbb{Z}} |\lambda_j|^2 < \infty, \quad \|\lambda\|^2_{L^2_{P_K}[-\pi,\pi]} = \frac{1}{\mu} \sum_{j \in \mathbb{Z}} |\lambda_j|^2.$$

Then,

$$(L_{P,K}\lambda)(t) = \big(L_{P,K} \sum_{j \in \mathbb{Z}} \lambda_j \phi_j\big)(t) = \sum_{j \in \mathbb{Z}} \lambda_j \mu_j \phi_j(t), \quad t \in [-\pi, \pi],$$

and, consequently,

$$\|L_{P,K}\lambda\|^2_{\mathcal{H}} = \sum_{j \in \mathbb{Z}} \frac{|\lambda_j|^2 \mu_j^2}{\mu_j} = \sum_{j \in \mathbb{Z}} \mu_j |\lambda_j|^2.$$

Also, for arbitrary $f \in \mathcal{H}$ as in (8.1) and (8.2), we have

$$\|f - L_{P,K}\lambda\|^2_{\mathcal{H}} = \big\|\sum_{j \in \mathbb{Z}} (c_j - \lambda_j \mu_j) \phi_j \big\|^2_{\mathcal{H}} = \sum_{j \in \mathbb{Z}} \frac{|c_j - \lambda_j \mu_j|^2}{\mu_j}.$$

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of points in $[-\pi,\pi]$. By Theorem 47 and taking into account of the inequality (7.6), for any $N\in\mathbb{N}$ and $\delta>0$ we have

$$\begin{aligned} P^N\big(\|f-\pi_x^N f\|_{\mathcal{H}}\geq\delta\big) &\leq P^N\big(\|f-\frac{1}{N}\sum_{n=1}^{N}\lambda(x_n)K_{x_n}\|_{\mathcal{H}}\geq\delta\big) \quad (8.4)\\ &\leq \frac{1}{\delta^2}\sum_{j\in\mathbb{Z}}\frac{|c_j-\lambda_j\mu_j|^2}{\mu_j}+\frac{1}{N\delta^2}\Big(\sum_{j\in\mathbb{Z}}(\mu-\mu_j)|\lambda_j|^2\Big) \end{aligned}$$

On the other hand, we observe that in the inequality (8.4) the left hand side does not depend on λ and hence, for any $\epsilon>0$ there exists $\lambda\in L^2_{P_K}[-\pi,\pi]$ such that

$$P^N\big(\|f-\pi_x^N f\|_{\mathcal{H}}\geq\delta\big)<\frac{\epsilon}{2}+\frac{1}{N\delta^2}\Big(\sum_{j\in\mathbb{Z}}(\mu-\mu_j)|\lambda_j|^2\Big),$$

and then, for sufficiently large N we get

$$P^N\big(\|f-\pi_x^N f\|_{\mathcal{H}}\geq\delta\big)<\epsilon.$$

In particular, if $f \in \mathcal{H}_P$, that is, the inequality (8.2) is replaced by the stronger one

$$\sum_{j\in\mathbb{Z}} \frac{|c_j|^2}{\mu_j^2} < \infty,$$

we can choose $\lambda_j = c_j/\mu_j$, $j \in \mathbb{Z}$, and we have $\lambda \in L^2_{P_K}[-\pi, \pi]$, hence

$$P^N\big(\|f - \pi_x^N f\|_{\mathcal{H}} \geq \delta\big) \leq \frac{1}{N\delta^2}\left(\sum_{j\in\mathbb{Z}} \frac{(\mu - \mu_j)|c_j|^2}{\mu_j^2}\right).$$

For example, this is the case for $f = \phi_k$ for some $k \in \mathbb{Z}$, hence $c_j = \delta_{j,k}$, $j \in \mathbb{Z}$, and letting $\lambda = \phi_k/\mu_k$, hence $\lambda_j = \delta_{j,k}/\mu_j$, $j \in \mathbb{Z}$, we have $f = L_{P,K}\lambda$ and hence,

$$P^N(\|\phi_k - \pi_x^N \phi_k\| \geq \delta) \leq \frac{1}{N\delta^2\mu_k^2} \sum_{\mathbb{Z} \ni j \neq k} \mu_j.$$

This shows that, the larger μ_k is, the faster ϕ_k will be approximated but, since $\mu_j \underset{j}{\to} 0$, ϕ_js cannot be approximated uniformly, in the sense that there does not exist a single N to make each $\left\|\phi_j - \pi_x^N \phi_j\right\|_{\mathcal{H}}$ bounded by the same δ with the same probability η.

This analysis can be applied more generally to kernels that admit an expansion analogous to (8.3) under basis functions $(\phi_j)_j$ which constitute a total orthonormal set in $L^2(X; P_K)$, e.g. as guaranteed by Mercer's Theorem [15, Theorem 2.30].

Example: The Hardy space $H^2(\mathbb{D})$

We consider the open unit disc in the complex plane $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ and the Szegö kernel

$$K(z,\zeta) = \frac{1}{1 - z\overline{\zeta}} = \sum_{n=0}^{\infty} z^n \overline{\zeta}^n, \quad z, \zeta \in \mathbb{D}, \tag{8.5}$$

where the series converges absolutely and uniformly on any compact subset of $\mathbb{D}$. The RKHS associated to K is the Hardy space $H^2(\mathbb{D})$ of all functions $f : \mathbb{D} \to \mathbb{C}$ that are holomorphic in $\mathbb{D}$ with power series expansion

$$f(z) = \sum_{n=0}^{\infty} f_n z^n, \tag{8.6}$$

such that the coefficients sequence $(f_n)_n$ is in $\ell^2_{\mathbb{C}}(\mathbb{N}_0)$.

The inner product in $H^2(\mathbb{D})$ is

$$\langle \sum_{n=0}^{\infty} f_n z^n, \sum_{n=0}^{\infty} g_n z^n \rangle = \sum_{n=0}^{\infty} f_n \overline{g_n},$$

with norm

$$\| \sum_{n=0}^{\infty} f_n z^n \|^2 = \sum_{n=0}^{\infty} |f_n|^2.$$

For each $\zeta \in \mathbb{D}$ we have

$$\|K_\zeta\| = \Big(\sum_{n=0}^{\infty} |\zeta|^{2n} \Big)^{1/2} = \frac{1}{\sqrt{1 - |\zeta|^2}},$$

hence the kernel K is unbounded.

We consider P the normalized Lebesgue measure on $\mathbb{D}$, that is, for $z = x + \mathrm{i}y = r\mathrm{e}^{\mathrm{i}\theta}$ we have

$$\mathrm{d}P(z) = \frac{1}{\pi}\,\mathrm{d}A(x,y) = \frac{r}{\pi}\,\mathrm{d}\theta\,\mathrm{d}r,$$

hence,

$$\mathrm{d}P_K(z) = \frac{r}{\pi(1-r^2)}\,\mathrm{d}\theta\,\mathrm{d}r.$$

Then, $L^2(\mathbb{D}; P_K)$ is contractively embedded in $L^2(\mathbb{D}; P)$.

Further on, in view of Proposition 46 and (8.5), for any $z, \zeta \in \mathbb{D}$ we have

$$K_P(z,\zeta) = \frac{1}{\pi}\int_0^1\int_0^{2\pi} \frac{r(1-r^2)}{(1-zr\mathrm{e}^{-\mathrm{i}\theta})(1-\overline{\zeta}r\mathrm{e}^{\mathrm{i}\theta})}\,\mathrm{d}\theta\,\mathrm{d}r \tag{8.7}$$

$$= \frac{1}{\pi}\int_0^1\int_0^{2\pi} \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}(1-r^2)r^{n+k+1}\mathrm{e}^{\mathrm{i}(n-k)\theta} z^n\overline{\zeta}^k\,\mathrm{d}\theta\,\mathrm{d}r$$

which, by using twice the Bounded Convergence Theorem for the Lebesgue measure, equals

$$= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\frac{1}{\pi}\int_0^1\int_0^{2\pi}(1-r^2)r^{n+k+1}\mathrm{e}^{\mathrm{i}(n-k)\theta}\,\mathrm{d}\theta\,\mathrm{d}r z^n\overline{\zeta}^k$$

$$= \sum_{n=0}^{\infty} 4\int_0^1 (1-r^2)r^{2n+1}\,\mathrm{d}r z^n\overline{\zeta}^n$$

$$= \sum_{n=0}^{\infty}\frac{z^n\overline{\zeta}^n}{(n+1)(n+2)}.$$

This shows that the RKHS $H^2_P(\mathbb{D})$ induced by K_P consists of all functions h that are holomorphic in $\mathbb{D}$ with power series

In order to calculate the operator $L_{P,K}\colon L^2(\mathbb{D}; P_K) \to H^2(\mathbb{D})$, let $\lambda \in L^2(\mathbb{D}; P_K)$ be arbitrary, that is, λ is a complex valued measurable function on $\mathbb{D}$ such that

$$\|\lambda\|^2_{L^2(\mathbb{D};P_K)} = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \frac{|\lambda(r\mathrm{e}^{\mathrm{i}\theta})|^2 r}{1-r^2}\,\mathrm{d}\theta\,\mathrm{d}r < \infty. \tag{8.8}$$

Then, in view of Proposition 39, we have

$$\begin{aligned}(L_{P,K}\lambda)(z) &= \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \lambda(r\mathrm{e}^{\mathrm{i}\theta}) K(z, r\mathrm{e}^{\mathrm{i}\theta}) r\,\mathrm{d}\theta\,\mathrm{d}r \\ &= \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \lambda(r\mathrm{e}^{\mathrm{i}\theta}) \sum_{n=0}^{\infty} z^n r^{n+1} \mathrm{e}^{-\mathrm{i}n\theta}\,\mathrm{d}\theta\,\mathrm{d}r\end{aligned} \tag{8.9}$$

which, by the Bounded Convergence Theorem, equals

$$= \sum_{n=0}^{\infty} \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \lambda(r\mathrm{e}^{\mathrm{i}\theta}) r^{n+1} \mathrm{e}^{-\mathrm{i}n\theta}\,\mathrm{d}\theta\,\mathrm{d}r z^n = \sum_{n=0}^{\infty} \lambda_n z^n,$$

where, for each integer $n \geq 0$ we denote

$$\lambda_n = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} \lambda(r\mathrm{e}^{\mathrm{i}\theta}) r^{n+1} \mathrm{e}^{-\mathrm{i}n\theta}\,\mathrm{d}\theta\,\mathrm{d}r. \tag{8.10}$$

Observing that, letting $\phi_n(z) := \sqrt{n+1}z^n$, for all integer $n \geq 0$ and $z \in \mathbb{D}$, the set $\{\phi_n\}_{n\geq 0}$ is orthonormal in $L^2(\mathbb{D}; P)$, it follows that $\lambda_n = \langle \lambda, \phi_n \rangle_{L^2(\mathbb{D};P)}$ for all integer $n \geq 0$ and, hence, $(\lambda_n)_{n\geq 0}$ is the weighted sequence of Fourier coefficients of λ with respect to the system of orthonormal functions $\{\phi_n\}_{n\geq 0}$ in $L^2(\mathbb{D}; P)$. On the other hand, since $L^2(\mathbb{D}; P_K)$ is contractively embedded in $L^2(\mathbb{D}; P)$, this shows that $L_{P,K}$ is the restriction to $L^2(\mathbb{D}; P_K)$ of a Bergman type weighted projection of $L^2(\mathbb{D}; P)$ onto a subspace of the Hardy space $H^2(\mathbb{D})$, that happens to be exactly $H^2_P(\mathbb{D})$.

Finally, let $f \in H^2(\mathbb{D})$ with power series representation as in (8.6) and let $\lambda \in L^2(\mathbb{D}; P_K)$ with norm given as in (8.8). Then, by Theorem 47 and taking into account of the inequality (7.6), for any $N \in \mathbb{N}$ and $\delta > 0$ we have

$$P^N\big(\|f - \pi_{\mathrm{z}}^N f\|_{H^2(\mathbb{D})} \geq \delta\big) \leq P^N\big(\|f - \frac{1}{N}\sum_{i=1}^{N} \lambda(z_i) K_{z_i}\|_{H^2(\mathbb{D})} \geq \delta\big) \tag{8.11}$$

$$\leq \frac{1}{\delta^2}\sum_{n=0}^{\infty} |f_n - \lambda_n|^2 + \frac{1}{N\delta^2}\big(\|\lambda\|^2_{L^2(\mathbb{D};P_K)} - \sum_{n=0}^{\infty} |\lambda$$

where $\mathrm{z} = (z_i)_{i \in \mathbb{N}}$ denotes an arbitrary sequence of points in $\mathbb{D}$ and π_{z}^N denotes the projection of $H^2(\mathbb{D})$ onto $\operatorname{span}\{K_{z_i} \mid i = 1, \ldots, N\}$.

By exploiting the fact that the left hand side in (8.11) does not depend on λ and the density of $H^2_P(\mathbb{D})$ in $H^2(\mathbb{D})$, for any $\varepsilon > 0$ there exists $\lambda \in L^2(\mathbb{D}; P_K)$ such that

$$P^N\left(\|f - \pi_z^N f\|_{H^2(\mathbb{D})} \geq \delta\right) \leq \frac{\varepsilon}{2} + \frac{1}{N\delta^2}\big(\|\lambda\|^2_{L^2(\mathbb{D};P_K)} - \sum_{n=0}^{\infty} |\lambda_n|^2\big),$$

and hence, for N sufficiently large, we have

$$P^N\left(\|f - \pi_z^N f\|_{H^2(\mathbb{D})} \geq \delta\right) \leq \varepsilon.$$

Let us consider now the special case when the function $f \in H_P^2(\mathbb{D})$, that is, with respect to the representation as in (8.6), we have the stronger condition

$$\sum_{n=0}^{\infty}(n+1)(n+2)|f_n|^2 < \infty.$$

In this case, letting

$$\lambda(z) := \sum_{n=0}^{\infty}(n+1)(n+2)(1-|z|^2)f_n z^n, \quad z \in \mathbb{D},$$

calculations similar to (8.7) and (8.9) show that

$$\frac{1}{\pi}\int_0^1\int_0^{2\pi}\frac{|\lambda(r\mathrm{e}^{\mathrm{i}\theta})|^2 r}{1-r^2}\,\mathrm{d}\theta\,\mathrm{d}r = \sum_{n=0}^{\infty}(n+1)(n+2)|f_n|^2 < \infty.$$

Then $\lambda \in L^2(\mathbb{D}; P_K)$, and

$$(L_{P,K}\lambda)(z) = \frac{1}{\pi}\int_0^\infty \int_0^{2\pi} \lambda(r\mathrm{e}^{\mathrm{i}\theta})K(z, r\mathrm{e}^{\mathrm{i}\theta})r \,\mathrm{d}\theta \,\mathrm{d}r = \sum_{n=0}^\infty f_n z^n = f(z), \quad z$$

hence, the first term in the right hand side of (8.11) vanishes and we get

$$P^N\left(\|f - \pi_z^N f\|_{H^2(\mathbb{D})} \geq \delta\right) \leq \frac{1}{N\delta^2}\sum_{n=0}^\infty (n^2+3n+1)|f_n|^2.$$

For example, if $f(z) = z^n$ for some integer $n \geq 0$, then

$$P^N\left(\|f - \pi_z^N f\|_{H^2(\mathbb{D})} \geq \delta\right) \leq \frac{n^2+3n+1}{N\delta^2},$$

showing that better approximations are obtained for smaller n than for bigger n.

N. Aronszajn, La théorie générale des noyaux reproduisants et ses applications, Premiére Partie, *Proc. Cambridge Philos. Soc.* **39**(1944), 133–153.

N. Aronszajn, Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, **68**(1950), 337–404.

D.L. Cohn, *Measure Theory*, Birkhäuser, New York 2013.

F. Cucker, S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39.1**(2002), 1–49.

J. Diestel, J.J. Uhl, Jr. *Vector Measures*, Mathematical Surveys and Monographs, No. 15, Amer. Math. Soc. 1977.

J.D. Eldredge, A. Leonard, T. Colonius, A general deterministic treatment of derivatives in particle methods, *Journal of Computational Physics*, **180**(2002), 686–709.

F. Girosi, Approximation error bounds that use VC bounds, in *Proc. International Conference on Artificial Neural*

Networks, F. Fogelman-Soulied and P. Gallinari (eds), pp. 295–302, Paris 1995.

F. Girosi, An equivalence between sparse approximation and support vector machines, *Neural Computation*, **10**(1998), 1455–1480.

J.H. Irving, J.G. Kirkwood, The statistical mechanical theory of transport processes, IV. The equations of hydrodynamics, *Journal of Chemical Physics*, **18**(1950), 817–829.

M.A. Kon, L.A. Raphael, Approximating functions in reproducing kernel Hilbert spaces via statistical learning theory, *Wavelets and Splines: Athens 2005*, pp. 271–286, *Mod. Methods Math.*, Nashboro Press, Brentwood, TN, 2006.

M.A. Kon, L.A. Raphael, D.A. Williams, Extending Girosi's approximation estimates for functions in Sobolev spaces via statistical learning theory, *J. Anal. Appl.* **3**(2005), 67–90.

H. Körezlioğlu, Reproducing kernels in separable Hilbert spaces, *Pacific J. Math.* **25**(1968), 305–314.

V.I. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, Cambridge 2016.

S. Reboux, B. Schrader; I.F. Sbalzarini, A self-organizing Lagrangian particle method for adaptive-resolution advection-diffusion simulations. *Journal of Computational Physics*, **231** (2012), 3623–3646.

S. Saitoh, Y. Sawano, *Theory of Reproducing Kernels and Applications*, Springer, Singapore 2016.

L. Schwartz, Sous espace Hilbertiens déspaces vectoriel topologiques et noyaux associés (noyaux reproduisants), *J. Analyse Math.* **13**(1964), 115–256.

A. Smola, A. Gretton, L. Song, B. Schölkopf, A Hilbert space embedding for distributions, in *International Conference on Algorithmic Learning Theory*, (2007), 13–31.

- V.N. VAPNIK, *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1998.
- V.N. VAPNIK, *The Nature of Statistical Learning Theory*, 2nd ed., Springer, 2000.
- V.N. VAPNIK, A.YA. CHERVONENKIS, *Theory of Pattern Recognition. Statistical Problems of Learning* [Russian], Izdat. "Nauka", Moscow, 1974.