Assortment of facts about RKHS

George Wynne

Imperial College London

Abstract

This is a set of notes outlining a collection of facts about Reproducing Kernel Hilbert Spaces that I find helpful to know. The content is sourced from a number of texts and I give reference to all sources, although they might not be to the very original papers from which the results are stated.

1 Introduction

Definition 1. *Let $\mathcal{X} \neq \emptyset$ and $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. We say k is positive definite is $\forall n \in \mathbb{N}\, \forall a_1, \dots, a_n \in \mathbb{R}^n$ and $\forall x_1, \dots, x_n \in \mathcal{X}$*

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Definition 2. *Let $\mathcal{X} \neq \emptyset$ and $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. We say k is a kernel if is it is symmetric and positive definite.*

Definition 3. *A Hilbert space of functions $\mathcal{H}$ over a non-empty set $\mathcal{X}$ is a Reproducing Kernel Hilbert Space if for every $x \in \mathcal{X}$ the evalution operator δ_x is continuous.*

Theorem 1. *A Hilbert space $\mathcal{H}$ of functions over a non-empty set $\mathcal{X}$ is an RKHS if and only if there exists a kernel k, called the reproducing kernel of $\mathcal{H}$ such that*

1. $k(\cdot, x) \in \mathcal{H}\, \forall x \in \mathcal{X}$
2. $\langle f, k(\cdot, x) \rangle = f(x)\, \forall f \in \mathcal{H}\, \forall x \in \mathcal{X}$

Proof. Suppose $\mathcal{H}$ is an RKHS. Then δ_x is continuous for every $x \in \mathcal{X}$ so by Riesz representation theorem there exists $\phi_x \in \mathcal{H}$ such that $\delta_x(f) = \langle f, \phi_x \rangle$. Define $k(x, y) = \langle \phi_x, \phi_y \rangle$ then clearly k satisfies the two properties above and it is left to the reader to see that k is a kernel. Now suppose there exists such a kernel k. Then by Cauchy-Schwartz $|\delta_x(f)| = |\langle f, k(\cdot, x) \rangle| \leq \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}}$ hence δ_x is continuous. $\square$

Theorem 2. *[Berlinet and Thomas-Agnan, 2004, Theorem 3] For every kernel there exists a unique RKHS with k as its reproducing kernel.*

2 Representation of RKHS

This section shall contain results which are the most general and abstract. The center piece shall be Mercer's theorem which underpins the main decription of an RKHS as a Hilbert space of functions which have a prescribed decay of basis coefficients with respect to a prescribed basis, both the decay rate and basis dictated by the kernel. Then we shall see how the RKHS can be viwed as the range of an integral operator associated with the kernel. Finally a description is given of the RKHS through pointwise evaluations. Most of the following results are from [Steinwart and Christmann, 2008, Section 4.5] which describes the most common, basic form of Mercer's theorem, and we refer the reader to [Steinwart and Scovel, 2012] which has a far more general analysis and highlights the importance of the support of the underlying measure μ that will feature in the following results.

2.1 Mercer's Theorem

We regurgitate [Steinwart and Christmann, 2008, Section 4.5]. Given a measurable space $\mathcal{X}$ and μ a σ-finite measure on $\mathcal{X}$ let k be a kernel on $\mathcal{X}$ with $\int_{\mathcal{X}} k(x,x)^2 d\mu(x) < \infty$. Define the integral operator $T_k \colon L^2(\mathcal{X},\mu) \to L^2(\mathcal{X},\mu)$ then T_k is compact, self-adjoint and positive so we by the spectral theorem we have a countable orthonormal system in $L^2(\mathcal{X},\mu)$ of eigenfunctions of T_k which we denote $(e_n)_{n\in\mathbb{N}}$ and the ordered sequence of corresponding positive eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ such that for $f \in L^2(\mathcal{X},\mu)$.

$$T_k f = \sum_{n=1}^{\infty} \lambda_n \langle f, e_n \rangle e_n$$

The spectral theorem only tells us that the $(e_n)_{n\in\mathbb{N}}$ form an orthonormal system but not always a basis, to get this extra informtion we need an assumption on μ which is used in Theorem 3 and Theorem 4. This is also discussed in [Cucker and Zhou, 2007, Chapter 4] which goes into a helpful level of detail regarding the technicalities. In short, $(e_n)_{n\in\mathbb{N}}$ only forms a basis if T_k is injective [Steinwart and Scovel, 2012, Proof Theorem 3.1]. A sufficient condition to ensure this is that k is strictly integrally positive definite.

Theorem 3. *[Steinwart and Christmann, 2008, Theorem 4.49] Let $\mathcal{X}$ be a compact metric space, k a continuous kernel, μ a finite Borel measure with supp(μ) $= \mathcal{X}$. Then for $(e_n)_{n\in\mathbb{N}}$ and $(\lambda_n)_{n\in\mathbb{N}}$ as above we have for $x, x' \in \mathcal{X}$*

$$k(x,x') = \sum_{n=1}^{\infty} \lambda_n e_n(x) e_n(x')$$

where the convergence is absolute and uniform.

Theorem 4. *[Steinwart and Christmann, 2008, Theorem 4.51] With the assumptions of Theorem 3 we have that*

$$\mathcal{H}_k = \left\{ \sum_{n=1}^{\infty} a_n \sqrt{\lambda_n} e_n \colon a \in l_2(\mathbb{N}) \right\}$$

equipped with the inner product

$$\langle f, g \rangle = \sum_{n=1}^{\infty} a_n b_n$$

where $f = \sum_{n=1}^{\infty} a_n \sqrt{\lambda_n} e_n$, $f = \sum_{n=1}^{\infty} b_n \sqrt{\lambda_n} e_n$, is the RKHS of k. Furthermore $T_k^{\frac{1}{2}} \colon L^2(\mathcal{X},\mu) \to \mathcal{H}_k$ is an isometric isomorphism.

Note this theorem tells us that $(\sqrt{\lambda_n}e_n)_{n\in\mathbb{N}}$ is an orthonomal basis for $\mathcal{H}_k$. This result is made in far greater generality in [Steinwart and Scovel, 2012] by removing the assumption that $\mathcal{X}$ is compact and further investigating the effect of μ in the representation of the kernel.

Theorem 4 shows us that the RKHS is made of weighted sums of the eigenbasis that arises from the kernel, with the decay of the coefficients determined by the decay of the eigenvalues of the integral oeprator induced by the kernel. The faster the decay of the eigenvalues, the faster the coefficients a_n of the eigenbasis in the expansion $f = \sum_{n=1}^{\infty} a_n e_n$ have to decay since they must satisfy $\frac{a_n}{\sqrt{\lambda_n}} \in l_2(\mathbb{N})$, so the a_n must roughly decrease at least as fast as the $\sqrt{\lambda_n}$. Additionally Theorem 4 reveals a relationship between $L^2(\mathcal{X},\mu)$ and the RKHS, namely the RKHS is the image of the operator $T_k^{\frac{1}{2}}$ which acts as a smoothing operator.

2.2 Mercer's Theorem Examples

Explicit Mercer expansions are somewhat hard to come by but there are asymptotic rates available for the eigenvalues for kernels of finite smoothness via n-width computations [Santin and Schaback, 2016, Section 5] which can aid some computations. Before we discuss some explicit examples note that the RKHS $\mathcal{H}_k$ does not depend on the μ being used, only on k, instead it is the basis used which changes. For a further discussion about the impact of the relationship between the RKHS and μ see the Final Remarks section of the Arxiv version of [Steinwart, 2018].

The largest list of Mercer expansions the author has seen is [Fasshauer and McCourt, 2014, Appendix A] which will not be fully derived again here since the list is so long. Included is the squared exponential kernel over $\mathbb{R}$ with Gaussian weight function, the exponential kernel over $[0,\infty)$ with exponential weight function, exponential kernel over an interval $[-L,L]$ and $[0,1]$ which both involve solving additional equations to get the parameters of the eigenfunctions, Brownian motion kernel over $[0,1]$, Brownian bridge kernel over $[0,1]$ and iterated Brownian bridge kernel over $[0,1]$.

2.3 Representation via sum of anchored kernels

The following representation of an RKHS is perhaps the most general construction since it makes no restrictions on the underlying space, but the statement has a completion step which obsfucates what functions are actually contained in the RKHS. The proof can be found in any book or reasonable set of lecture notes about RKHS so if you don't like the presentation in the referenced proof then fear not because you should be able to quickly find a different reference.

Theorem 5. *[Cucker and Zhou, 2007, Theorem 2.9] Let $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a kernel then $\mathcal{H}_k$ is the completion of the set*

$$\mathcal{H}_{pre} = \left\{\sum_{n=1}^{N} a_n k(\cdot,x_n)\colon N\in\mathbb{N}, (a_n)_{n=1}^N\subset\mathbb{R}, (x_n)_{n=1}^N\subset\mathcal{X}\right\}$$

with respect to the inner product

$$\langle f,g\rangle = \sum_{n=1}^{N}\sum_{m=1}^{M} a_n b_m k(x_n,y_m)$$

for $f=\sum_{n=1}^N a_n k(\cdot,x_n)$ and $g=\sum_{m=1}^M b_m k(\cdot,y_m)$.

2.4 Representation via pointwise evaluation functionals

Since $k(\cdot, x)$ is the Riesz representor of the linear operator δ_x in $\mathcal{H}_k$ a different view of Theorem 5 can be employed which involves linear combinations of point evaluation functionals rather than linear combinations of anchored kernels. This other representation seems to be not well known outside of the scattered data approximation literature and facilitates the proof of very interesting "inverse" theorems for RKHS [Schaback and Wendland, 2002]. A discussion is also given in [Wendland, 2004, Chapter 10.4] which also includes conditionally positive definite functions. Since these types of functions are often not considered in kernel methods or machine learning the proof is reproduced here without the generality to hold for such functions.

Theorem 6. *[Wendland, 2004, Theorem 10.22] Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous kernel. Define the space*

$$\Lambda = \left\{ \lambda = \sum_{n=1}^{N} a_n \delta_{x_n} : N \in \mathbb{N}, (a_n)_{n=1}^N \subset \mathbb{R}, (x_n)_{n=1}^N \subset \mathcal{X} \right\}$$

with the inner product

$$\langle \lambda, \gamma \rangle = \sum_{n=1}^{N} \sum_{m=1}^{M} a_n b_m k(x_n, y_m)$$

for $\lambda = \sum_{n=1}^N a_n \delta_{x_n}$ and $\gamma = \sum_{m=1}^M b_m \delta_{y_m}$ then the space

$$\mathcal{G} = \{ f \in C(\mathcal{X})\colon \exists C_f > 0 \text{ such that } |\lambda(f)| \leq C_f \|\lambda\|_\Lambda \; \forall \lambda \in \Lambda \}$$

with the norm

$$\|f\|_{\mathcal{G}} = \sup_{\|\lambda\|_\Lambda \leq 1} |\lambda(f)|$$

is equal to the space $\mathcal{H}_k$ and the norms are equal.

Proof. Given $\lambda = \sum_{n=1}^N a_n \delta_{x_n} \in \Lambda$ we will use the notation $\lambda^x k(\cdot, x)$ to denote $\sum_{n=1}^N a_n k(\cdot, x_n)$ i.e. λ^x is the action of λ on the second coordinate of k. With this notation we have by the reproducing property

$$\lambda(f) = \sum_{n=1}^{N} a_n f(x_n) = \sum_{n=1}^{N} a_n \langle f, k(\cdot, x_n) \rangle = \langle f, \lambda^x k(\cdot, x) \rangle$$

so by Cauchy-Schwartz

$$|\lambda(f)| \leq |\langle f, \lambda^x k(\cdot, x) \rangle| \leq \|f\|_{\mathcal{H}} \|\lambda^x k(\cdot, x)\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} \|\lambda\|_{\mathcal{G}}$$

by the definition of the norm on $\mathcal{G}$, hence setting $C_f = \|f\|_{\mathcal{H}}$ shows $f \in \mathcal{G}$ and $\|f\|_{\mathcal{G}} \leq \|\mathcal{X}\| f_{\mathcal{H}}$. We know that $f \in C(\mathcal{X})$ since we assumed k is continuous.

Now assume $f \in \mathcal{G}$ and note that every element of $\mathcal{H}_{pre}$ can be expressed uniquely as $\lambda^x k(\cdot, x)$ for some $\lambda \in \Lambda$. The function f induces an operator $F_f\colon \mathcal{H}_{pre} \to \mathbb{R}$ defined as

$$F_f(\lambda^x k(\cdot, x)) = \lambda(f)$$

Since $\mathcal{H}_{pre}$ is dence in $\mathcal{H}_k$ we know we can extend F_f to a linear operator on $\mathcal{H}$ and by Riesz there exists $h \in \mathcal{H}$ such that $F_f(g) = \langle g, h \rangle_{\mathcal{H}_k}\ \forall g \in \mathcal{H}_k$. In particular for any $x \in \mathcal{X}$ we can take δ^x and observe that

$$F_f(k(\cdot, x)) = f(x) = \langle k(\cdot, x), h \rangle = h(x)$$

where the first equality is by definition of F_f the second is by Riesz and the third by reproducing property. Since x was arbitrary we can conclude $f = h \in \mathcal{H}_k$.

To conclude that the norms are equal note that since $\mathcal{H}_{pre}$ is dense in $\mathcal{H}_k$ we can find a sequence λ_n in Λ such that $\lambda_n^x k(\cdot, x)$ converges to f in $\mathcal{H}_k$, this means that $\lambda_n(f) = \langle f, \lambda_n^x k(\cdot, x) \rangle \to \|f\|_{\mathcal{H}_k}^2$ and $\|\lambda_n\|_{\mathcal{G}} = \|\lambda_n^x k(\cdot, x)\|_{\mathcal{H}_k} \to \|f\|_{\mathcal{H}_k}$. All this allows use to conlude that

$$\|f\|_{\mathcal{G}} \geq \lim_{n \to \infty} \frac{|\lambda_n(f)|}{\|\lambda_n\|_{\mathcal{G}}} = \frac{\|f\|_{\mathcal{H}_k}^2}{\|f\|_{\mathcal{H}_k}} = \|f\|_{\mathcal{H}_k}$$

□

The main moral of this proof is that we leverage the proof of Theorem 5 but use the Riesz representation theorem to have linear combinations of pointwise evaluation rather than linear combinations of sums of kernels. Since $\mathcal{H}_k$ is a Hilbert space it is the same as its dual and to be in the dual of $\mathcal{H}_k$ you need to have bounded operator norm over the elements of $\mathcal{H}_k$. The norm of $\mathcal{G}$ is an operator norm over a dense subset of $\mathcal{H}_k$ (using the natural identification of Λ and $\mathcal{H}_{pre}$). So a finite operator norm on the dense set gives a finite operator norm over $\mathcal{H}_k$ so we can conclude the function is in $\mathcal{H}_k$. Later in the present text we will explore very interesting applications of this pointwise representation of the RKHS and use it to show otherwise difficult to prove results.

2.5 Representation via Fourier transform

This final representation of the RKHS is perhaps the most helpful but imposes more requirements on the kernel. It requires that the kernel can be written as $k(x, y) = \phi_k(x - y)$ for some function ϕ_k and then the RKHS is expressed in terms of the Fourier transform of. ϕ_k. For the most used kernels we either have explciit expressions for their Fourier transform e.g. squared-exponential, exponential and Matérn, or we have asymptotic rates e.g. Wendland kernel so the next result is very helpful.

Theorem 7. *[Wendland, 2004, Theorem 10.12] Suppose* $k\colon \mathbb{R}^d \times \mathbb{R}^d$ *is a kernel with* $k(x, y) = \phi_k(x - y)$ *with* $\phi_k \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ *then*

$$\mathcal{H}_k = \left\{ f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \colon \hat{f}/\sqrt{\hat{\phi}_k} \in L^2(\mathbb{R}^d) \right\}$$

with inner product

$$\langle f, g \rangle_{\mathcal{H}_k} = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \frac{\hat{f}(x)\overline{\hat{g}(x)}}{\hat{\phi}_k(x)} dx$$

In particular if $\hat{\phi}_k$ has algebraic decay then we can immediately conclude that the RKHS over $\mathbb{R}^d$ is norm equivalent to a Sobolev space, this is the [Wendland, 2004, Corollary 10.13].

2.6 Representation via Fourier transform example

The most common example of an application of Theorem 7 is for the Matérn kernel. For a full derivation see [Kanagawa et al., 2018, Section 2.4].

2.7 Representation of squared exponential RKHS

This section is about the RKHS of the squared exponential, or Gaussian, kernel. This kernel is extremely popular in practise for a variety of reasons, partly due to strong mathematical theory and partly due to decades of convention. The problem of describing the RKHS has attracted a lot of attention with three papers released at similar times [Steinwart et al., 2006, van der Vaart and van Zanten, 2008, Minh, 2009] being the most well known and taking different approaches to the problem of describing the RKHS.

The first is from a Bayesian non-parametrics point of view and derives results related to contraction of Gaussian process regression which isn't the focus of this note. The second focusses on the insight that the Gaussian kernel has special restriction properties due to analyticity and any RKHS over a set with non-empty interior is the exact restriction of the kernel over $\mathbb{C}^d$, which can be used to deduce interesting properties. The third focusses on using the Weyl inner product to give the clearest description of the RKHS from which elementary computations can yield surprising properties about the space.

We focus on [Minh, 2009] since the results are stated the most easily and have a clear message. The statements uses multi-index notation where for $\alpha \in \mathbb{N}^d$ we set $|\alpha| = \sum_{n=1}^{d} \alpha_n$ and for $x \in \mathbb{R}^d$ we set $x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$

Theorem 8. *[Minh, 2009, Theorem 1] Let $\mathcal{X} \subset \mathbb{R}^d$ be any set with non-empty interior. Let $k(x,y) = \exp(-\frac{1}{2}\frac{\|x-y\|_2^2}{l^2})$ then*

$$\mathcal{H}_k(\mathcal{X}) = \left\{ e^{-\frac{\|x\|_2^2}{l^2}} \sum_{|\alpha|=0}^{\infty} w_\alpha x^\alpha \colon \|f\|_{\mathcal{H}_k}^2 = \sum_{n=0}^{\infty} \frac{n! l^{2n}}{2^n} \sum_{|\alpha|=n} \frac{w_\alpha^2}{C_\alpha^n} < \infty \right\}$$

where $C_\alpha^n = \frac{n!}{\alpha_1! \dots \alpha_d!}$. The inner product is given by

$$\langle f, g \rangle = \sum_{n=0}^{\infty} \frac{n! l^{2n}}{2^n} \sum_{|\alpha|=n} \frac{w_\alpha v_\alpha}{C_\alpha^n}$$

where $f = e^{-\|x\|_2^2 l^2} \sum_{|\alpha|=0}^{\infty} w_\alpha x^\alpha$ and $g = e^{-\frac{\|x\|_2^2}{l^2}} \sum_{|\alpha|=0}^{\infty} v_\alpha x^\alpha$ and an orthonormal basis for $\mathcal{H}_k(\mathcal{X})$ is

$$\phi_\alpha(x) = \sqrt{\frac{2^{|\alpha|} C_\alpha^{|\alpha|}}{l^{2|\alpha|} |\alpha|!}} e^{-\frac{\|x\|_2^2}{l^2}} x^\alpha$$

This shows that the squared exponential RKHS is composed of exponentially damped multinomials and functions which are not exponentially damped will have a high norm. Three more interesting results from [Minh, 2009] are presented next. All proofs are done in [Minh, 2009] via direct calculation using the above formula for the norm of the RKHS. Theorem 9 generalises [Steinwart et al., 2006, Corollary 3.9] by deducing that the RKHS doesn't contain polynomials. Theorem 10 can also be proven by substituting in the Fourier transform of the Gaussian kernel into Theorem 7 and using [Steinwart et al., 2006, Corollary 3.8].

Theorem 9. *[Minh, 2009, Theorem 2] Under the same assumptions as Theorem 8, $\mathcal{H}_k(\mathcal{X})$ does not contain any polynomial on $\mathcal{X}$, including non-zero constant functions.*

Theorem 10. *[Minh, 2009, Theorem 3] Under the same assumptions as Theorem 8 the function $e^{-\frac{\lambda\|x\|_2^2}{l^2}}$ is in $H_k(\mathcal{X})$ if and only if $0 < \lambda < 2$.*

Theorem 11. *[Minh, 2009, Theorem 4] For any $l > 0$ the RKHS of $e^{-\frac{\|x\|_2^2}{l^2}}$ over $\mathbb{R}^d$ is not a subset of $L^1(\mathbb{R}^d)$.*

2.8 RKHS characterisation via norm of interpolant

The next result [Schaback and Wendland, 2002, Theorem 5.1] is an interesting characterisation of functions in the RKHS which features the norm of the kernel interpolant. First recall that if you have a function $f \notin \mathcal{H}_k$ then $\|f\|_{\mathcal{H}_k} = \infty$, the intuition is then if we observe $\{f(x_i)\}_{n=1}^N$ for some collection of points $X = \{x_i\}_{n=1}^N$ and fit a kernel interpolant to f, denoted $I_X f$, then $I_X f \approx f$ therefore $\|I_X f\|_{\mathcal{H}_k} \approx \|f\|_{\mathcal{H}_k}$ meaning we would expect the RKHS norm of the kernel interpolant to grow. This if formalised in the following result.

Theorem 12. *[Schaback and Wendland, 2002, Theorem 5.1] Let k be a kernel and $\mathcal{H}_k$ be the RKHS. Then $f \in \mathcal{H}_k$ if and only if there exists a constant $C > 0$, which can depend on f, such that $\|I_X f\|_{\mathcal{H}_k} \leq C$ for any collection of finite points X where $I_X f$ is the kernel interpolant of f given the observations $\{f(x) \colon x \in X\}$*

This theorem is of practical importance in the scenario of approximating an unkown function using kernel interpolation. This is because theoretical bounds often involve the quantity $\|I_X f\|_{\mathcal{H}_k}$ which, by the above result, will grow as more data is observed if $f \notin \mathcal{H}_k$. For quantitative rates on the growth of this quantity see [Karvonen et al., 2020, Theorem 4.9].

3 Literature guide

The aim of this section is to provide an overview of common references for RKHS, both textbooks and large papers. I will give a shrot description of the contents and the flavour of results one would find in each source. The sources are not presented in any particular order and hints of how the references are related to each other are included so you should read all the decriptions to get the most out of this section!

3.1 Textbooks

[Steinwart and Christmann, 2008] studies statistical properties of Support Vector Machines (SVM) which are method of approximating functions and creating calssifiers using kernels. The RKHS is a vital part of the analysis since it describes the capacity of the model. Chapter 4 of this book is dedicated to RKHS. It describes the feature map view of kernels which is typical for the SVM literature. Of particular interest in this source is a thorough description of properties inheirited by functions in the RKHS from the kernel. A description of the Gauss RKHS is given along with a discussion of Mercer's theorem.

[Berlinet and Thomas-Agnan, 2004] discusses the relationship between kernel, stochastic processe and statistical estimators. Numerous examples of kernels and RKHS are given. Chapter 1 gives a large description of properties of RKHS. This source if of particular interest to those interested in Gaussian processes as it discusses the links between kernels methods and GPs.

[Wendland, 2004] provides a comprehensive overview of the field Scattered Data Approximation (SDA) which is the study of numerical approxiamtion methods using translation invariant kerenels, referred to as radial basis functions in the literature. The RKHS is referred to as the native space. This source provides numerous helpful results regarding the identification of RKHS and properties of common translation invariant kernels. For interest of numerical analysts is the treatment of approximation error bounds in terms of properties of the observed data points. These results are important for modelling functions with GPs due to the intimate relationship between kernel methods and Gaussian processes.

[Paulsen and Raghupathi, 2016] provides a concise collection of helpful results regarding mathematical properties of kernels and RKHS. The second half of this text then applies the theory to various domains. This text has a more classical definition, proposition, theorem layout than some of the other references due to its compact size. Additionally it places emphasis on some kernels that are studied more in pure maths than statistics and machine learning such as the Hardy kernel. These facts are rather interesting and are well presented to still read it if you don't have a pure math background!

[Fasshauer and McCourt, 2014] is a reference text which provides a modern presentation of SDA results and their relationship to more machine learning based topics. This book concerntrates on numerical implementation with a very large number of helpful examples and extensive reference list.

3.2 Large/Survey papers

[Aronszajn, 1950] provides perhaps the first full treatment of the theory of RKHS and contains numerous original results with clear proofs.

[Kanagawa et al., 2018] provides a very accessible survey of the relationship between kernel methods and GPs. In the first half of the paper a discussion of RKHS is given and for people interested in kernel methods more generally this paper is a must read. Topics include function approxiation, maximum mean discrepency, regularity of sample paths of GPs and Bayesian quadrature.

[Cucker and Smale, 2001] lays the foundations of statistical learning and describes the problem of function approximation in this context. The impact of RKHS is clearly given. This is a mathematical paper so might require some extra reading to understand.

[Schaback and Wendland, 2006] is a large survey paper of SDA results and their relationship to machine learning. I would recommend reading it before [Wendland, 2004] to understand the framework of these sort fo results. Emphasis is placed on the numerical analysis aspects of the methods.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68 (3):337–337, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(01):1–50, October 2001. doi: 10.1090/s0273-0979-01-00923-5. URL `https://doi.org/10.1090/s0273-0979-01-00923-5`.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory*. Cambridge University Press, 2007.

G. Fasshauer and M. McCourt. *Kernel-based Approximation Methods using MATLAB*, volume 19 of *Interdisciplinary Mathematical Sciences*. World Scientific, 2014.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018.

Toni Karvonen, George Wynne, Filip Tronarp, Chris J. Oates, and Simo Särkkä. Maximum likelihood estimation and uncertainty quantification for gaussian process approximation of deterministic functions, 2020.

Ha Quang Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2009.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016. doi: 10.1017/cbo9781316219232. URL `https://doi.org/10.1017/cbo9781316219232`.

G. Santin and R. Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

Robert Schaback and Holger Wendland. Inverse and saturation theorems for radial basis function interpolation. *Mathematics of Computation*, 71(238):669–681, 2002.

Robert Schaback and Holger Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numerica*, 15:543–639, May 2006. doi: 10.1017/s0962492906270016. URL `https://doi.org/10.1017/s0962492906270016`.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

Ingo Steinwart. Convergence types and rates in generic karhunen-loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395, 2018.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel hilbert spaces of gaussian priors. In *Institute of Mathematical Statistics Collections*, pages 200–222. Institute of Mathematical Statistics, 2008.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.