Reproducing Kernel Hilbert Spaces

Thomas Zhang

Contents

1 Overview of Important Properties

Let's say that we had some data that lie on non-linear manifolds such that linear classification isn't very effective. We can send the data to a higher dimension space such that the data now lie on linear manifolds in the higher dimension space. One of the broad motivations behind RKHS is to be able to analyze and interpret the embedded data. We now introduce some fundamental terminology.

Definition 1.1 ((Reproducing) Kernel) *Say that our data is in a set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is a kernel if it satisfies the following properties:*

1. *(Symmetric) For all $x, y \in \mathcal{X}$, $k(x, y) = k(y, x)$.*
2. *(Positive definite) For any $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the $n \times n$ Gram matrix K where $K_{ij} = k(x_i, x_j)$ is positive semi-definite (semi-definite iff data is linearly dependent). Equivalently: for all $v \in \mathbb{R}^n$, $v^\top K v \geq 0$.*

A kernel is furthermore a "reproducing" kernel of a space $\mathcal{H}$ if for all $x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$ and for any function $f \in \mathcal{H}$,

$$\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} = f(x).$$

Definition 1.2 (Canonical Feature Map) *Given set $\mathcal{X}$ that contains our data, we define the canonical feature map (a function that embeds the data to a higher-dimension space) to be*

$$\Phi : \mathcal{X} \to \mathcal{H}, \quad \Phi(x) = k(x, \cdot)$$

The order that we introduced these definitions is slightly deceptive: one can be characterized using the other and many things become tautologies. For example, we can actually construct a corresponding Hilbert space $\mathcal{H}$ given a reproducing kernel by taking the closure of all finite linear combinations of $k(x, \cdot)$:

$$\mathcal{H} := \overline{\Big\{ \sum_{i=1}^{m} c_i k(x_i, \cdot) : x_i \in \mathcal{X}, c_i \in \mathbb{F} \Big\}} \equiv \overline{\text{span}\big(k(x_i, \cdot),\, x_i \in \mathcal{X}\big)}.$$

The closure is necessary because span may not result in a closed linear subspace in infinite-dimensions. We further drive in the equivalence between feature maps and reproducing kernels: if $k(x, x) := \langle \phi(x), \phi(x) \rangle$ is a reproducing kernel, then

$$\begin{aligned} & k(x, y) \equiv \langle k(x, \cdot), k(\cdot, y) \rangle \\ \implies\ & k(x, x) \equiv \langle k(x, \cdot), k(\cdot, x) \rangle \\ \implies\ & \phi(x) \equiv k(x, \cdot), \end{aligned}$$

in other words, by defining a reproducing kernel from a given feature map, the feature map becomes the canonical feature map of the corresponding RKHS.

1.1 A Trio of Important Theorems

We have shown earlier than given a PSD (positive *symmetric* definite) kernel, we can construct a Hilbert space such that the kernel is a reproducing kernel. One might wonder, similar to the bijective relationship between feature maps and kernels, whether there is a bijective relationship between a PSD kernel and RKHS's. The following theorem establishes the bijectivity.

Theorem 1.3 (Moore-Aronszajn [1]) *Given a symmetric, positive definite kernel k on a set $\mathcal{X}$, there is a* unique *Hilbert space of functions on $\mathcal{X}$ for which k is a reproducing kernel.*

Proof: Most of the proof is by construction. We construct a Hilbert space $\mathcal{H}_0$ by taking the linear span of vectors $k(x, \cdot)$, where $x \in \mathcal{X}$. We define the inner product on $\mathcal{H}_0$ to be

$$\begin{aligned} \Big\langle \sum_{i=1}^{n} a_i k(x_i, \cdot), \sum_{j=1}^{m} b_i k(x_j, \cdot) \Big\rangle &= \sum_{i=1}^{n} a_i \Big\langle k(x_i, \cdot), \sum_{j=1}^{m} b_i k(x_j, \cdot) \Big\rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(x_i, x_j). \end{aligned}$$

We see that this inner product is well-defined and non-degenerate thanks to the symmetricity and positive-definiteness of $k(x, y)$. We then consider the completion of H_0; call it $\mathcal{H}$. Observe that even if $\mathcal{H}$ is not separable, we know that each element in $\mathcal{H}$ can be represented as a sum of at most countably many orthonormal basis elements (I can't recall the name of the theorem). Therefore, we can express any element of $\mathcal{H}$ in the form: $f(\cdot) = \sum_{i=1}^{\infty} c_i k(x_i, \cdot)$,

where $\sum_{i=1}^{n} a_i^2 k(x_i, x_i) < \infty$. All such f are well-defined thanks to the Cauchy-Schwarz inequality. We can verify that $k(x, y)$ is a reproducing kernel on $\mathcal{H}$:

$$\langle f(\cdot), k(x, \cdot)\rangle = \Big\langle \sum_{i=1}^{\infty} c_i k(x_i, \cdot), k(x, \cdot)\Big\rangle = \sum_{i=1}^{\infty} c_i k(x_i, x) = f(x).$$

To show that $\mathcal{H}$ is unique, We consider G where $k(x, y)$ is also a reproducing kernel. Observe that by our construction of $\mathcal{H}$, $G \supset \mathcal{H}$. Since $\mathcal{H}$ is complete, $\mathcal{H}$ must be a proper, closed linear subspace of G. We can therefore consider the decomposition $G = \mathcal{H} \oplus \mathcal{H}^{\perp}$. For any element in G we can write it as $g(\cdot) = f_{\mathcal{H}}(\cdot) + f_{\mathcal{H}^{\perp}}(\cdot)$. Observe that

$$\begin{aligned}\langle g(\cdot), k(x, \cdot)\rangle &= \langle f_{\mathcal{H}}(\cdot) + f_{\mathcal{H}^{\perp}}(\cdot), k(x, \cdot)\rangle \\ &= \langle f_{\mathcal{H}}(\cdot), k(x, \cdot)\rangle + \langle f_{\mathcal{H}^{\perp}}(\cdot), k(x, \cdot)\rangle \\ &= \langle f_{\mathcal{H}}(\cdot), k(x, \cdot)\rangle + 0 \\ &= f_{\mathcal{H}}(x).\end{aligned}$$

Observe that the only way for g to fulfill the reproducing property is if $f_{\mathcal{H}^{\perp}}(\cdot) = 0$, which implies that $\mathcal{H}^{\perp}$ is trivial. Therefore $G = \mathcal{H}$, establishing the uniqueness of $\mathcal{H}$ as the RKHS associated with the kernel $k(x, y)$. ∎

One benefit of embedding data into a Hilbert space is the fact that we have access to spectral theorems. Familiar to those who work with differential equations, given a PSD function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is jointly continuous on a compact domain $\mathcal{X}$, we can define a bounded linear (integral) operator on the Hilbert space $L^2(\mathcal{X})$: $T_k(f) := \int_{\mathcal{X}} k(\cdot, y) f(y) d\mu(y)$. Through this interpretation of the kernel function, we can characterize the corresponding RKHS and feature map through the eigen-decomposition of the integral operator. Mercer's theorem formalizes this notion:

Theorem 1.4 (Mercer) *Let $\mathcal{X}$ be a compact space equipped with a positive, finite Borel measure μ. Suppose $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a jointly continuous, positive definite, symmetric function. We define the integral operator $T_k(\cdot) \in B(L^2(\mathcal{X}), L^2(\mathcal{X}))$ the space of bounded linear operators $L^2(\mathcal{X}) \to L^2(\mathcal{X})$:*

$$T_k(f) = \int_{\mathcal{X}} k(\cdot, y) f(y)\, d\mu(y),$$

Then $T_k(f)$ can be diagonalized to yield a countable set of eigenvalues $\{\lambda_i\}$ and corresponding eigenvectors $\{\psi_i\}$ such that $\lim_{i\to\infty} \lambda_i \to 0$ and $\{\psi_i\} \subset L_2(\mathcal{X})$ form an orthonormal basis for $L_2(\mathcal{X})$. We can then write

$$\begin{aligned}k(x, y) &= \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y) \\ k(\cdot, \cdot) &= \sum_{i}^{\infty} \lambda_i (\psi_i \otimes \psi_i).\end{aligned}$$

Additionally, the corresponding RKHS can be expressed as

$$\mathcal{H} = \Big\{ f \in L_2(\mathcal{X}) : \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle^2}{\lambda_i} < \infty \Big\},$$

where the inner product is defined: $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle \langle g, \psi_i \rangle}{\lambda_i}$

Proof idea: Almost all the listed properties are a result of the extremely nice properties of the operator T_k. Observe that since $k(x, y)$ is symmetric kernel function, and $L^2(\mathcal{X})$ is separable, T_k is a self-adjoint Hilbert-Schmidt operator, which is in turn compact. The spectral theorem for self-adjoint compact operators tells us that the spectrum is real and accumulates only at 0, while the corresponding eigenvectors form an orthonormal basis for $L^2(\mathcal{X})$. Furthermore, since $k(x, y)$ is positive definite, the eigenvalues must all be positive. The decomposition of k is analogous to the diagonalization and representation of a positive definite matrix as a sum of projection operators, which is where we get:

$$k(\cdot, \cdot) = \sum_{i}^{\infty} \lambda_i (\psi_i \otimes \psi_i)$$

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y).$$

As for the representation of the RKHS: $\mathcal{H} = \Big\{ f \in L_2(\mathcal{X}) : \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle^2}{\lambda_i} < \infty \Big\}$, we essentially get this by re-defining our feature maps $\phi_i := \sqrt{\lambda_i} \psi_i$. ∎

One might wonder how blowing up the dimensionality of the data could ever be computationally sensible. While not the end-all-be-all solution to that problem, the following theorem at least reduces tells us the infinite-dimensional problem can be equivalently formulated in finite dimensions without sacrificing *any* accuracy.

Theorem 1.5 (Representer Theorem) *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel and $\mathcal{H}$ be its corresponding RKHS. Given sample data $(x_1, y_1), \ldots, (x_n, y_n) \subset \mathcal{X} \times \mathbb{R}$, a monotonically increasing function $g : [0, \infty) \to \mathbb{R}$, and any empirical loss function ℓ, we can express any*

$$f^* = \arg\min_{f \in \mathcal{H}} \left(\ell\big(f(x_1), \ldots, f(x_n)\big) + g(\|f\|_{\mathcal{H}}) \right)$$

as a finite sum

$$f^*(\cdot) = \sum_{i=1}^{n} c_i k(x_i, \cdot) \in \mathit{span}\big(k(x_1, \cdot), \ldots, k(x_n, \cdot)\big).$$

Proof: The proof follows from an appeal to orthogonality (praise Hilbert spaces). We consider the subspace $W \subset \mathcal{H}$ generated by the span of $k(x_i, \cdot)$. Since we have finitely many $k(x_i, \cdot)$, the subspace is closed. We may then decompose $\mathcal{H} = W \oplus W^{\perp}$ such that for any $f \in \mathcal{H}$, we can write it as

$$f(\cdot) = \sum_{i=1}^{n} k(x_i, \cdot) + r(\cdot), \quad r(\cdot) \in W^{\perp}.$$

Now we observe that since $\mathcal{H}$ is a RKHS,

$$\begin{aligned} f(x_j) &= \langle f(\cdot), k(x_j, \cdot)\rangle \\ &= \Big\langle \sum_{i=1}^{n} k(x_i, \cdot) + r(\cdot), k(x_j, \cdot)\Big\rangle \\ &= \Big\langle \sum_{i=1}^{n} k(x_i, \cdot), k(x_j, \cdot)\Big\rangle + \langle r(\cdot), k(x_j, \cdot)\rangle \\ &= \sum_{i=1}^{n} k(x_i, x_j) + 0 \quad \text{since } r(\cdot) \in W^{\perp}. \end{aligned}$$

Notice that we could've set the residual term r in $W^{\perp}$ to be anything and it would not have affected the value of f evaluated at any x_j. Keeping this in mind, we consider the second term $g(\|f\|_{\mathcal{H}})$. Using the Pythagorean identity, we can re-write this:

$$\begin{aligned} g(\|f\|_{\mathcal{H}}) &= g\bigg(\Big\| \sum_{i=1}^{n} k(x_i, \cdot) + r(\cdot) \Big\|_{\mathcal{H}} \bigg) \\ &= g\bigg(\Big\| \sum_{i=1}^{n} k(x_i, \cdot) \Big\|_{\mathcal{H}} + \|r(\cdot)\|_{\mathcal{H}} \bigg). \end{aligned}$$

Observe that since g is a strictly monotonically increasing function, we strictly decrease its value if we remove the residual term r. However, we must look toward the first term to make sure we don't somehow increase its value. Fortunately, we already established that the value of f evaluated at x_j is completely independent of r, so the value of the loss function is also independent of r. Therefore, our best bet is to simply set $r(\cdot) = 0$, which implies that the optimal function f^* is in the span of $k(x_i, \cdot)$, i.e.

$$f^*(\cdot) = \sum_{i=1}^{n} c_i k(x_i, \cdot) \in \text{span}\big(k(x_1, \cdot), \ldots, k(x_n, \cdot)\big). \quad \blacksquare$$

2 Universal Kernels

2.1 General Results

Everything in this section is derived from [3]. The motivation for universal kernels is straightforward: we know by Moore-Aronsazjn that with each kernel we can generate a unique RKHS, where elements in that RKHS can by definition be uniformly approximated by $k(x_i, \cdot)$, where x_i come from the input space $\mathcal{X}$. This is about as much as we can say in terms of the approximation power of a kernel when we must consider the whole input *space* (say $\mathbb{R}^n$). However, often times the actual inputs might not come from the entire space $\mathcal{X}$; in fact, given certain scenarios, the actual subset of the input space from which inputs are drawn might even be compact! One might now wonder: if we restrict our attention to a compact subset of the input space, say $Z \subset \mathcal{X}$, and consider the space of *all* continuous

functions on just that compact subset endowed with the sup norm $\|\cdot\|_\infty$, denoted $C(Z)$, now can our kernel approximate any element of $C(Z)$? More formally, we denote

$$K(Z) := \overline{span}(k(x,\cdot),\ x \in Z)$$

and we want to know if $K(Z)$ is dense in $C(Z)$. If it is dense, then we say the kernel $k(x,y)$ has the *universal approximating property*, and that $k(x,y)$ is a *universal kernel*.

Lemma 2.1 *If $K(Z)$ is dense in $C(Z)$, then in fact $K(Z) = C(Z)$.*

The lemma follows from the fact that $K(Z)$ is defined to be a closed linear subspace.

Now we dive in headfirst to the functional analysis-y part of the paper. Suppose we are given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a feature map $\Phi : \mathcal{X} \to \mathcal{H}$, and the corresponding Hilbert space $\mathcal{H}$. To make claims about $C(Z)$, we first have to understand its structure. I guess the natural starting place for that is to establish the dual space of $C(Z)$. Since Z is compact, it is trivially locally compact Hausdorff and every continuous function over Z also trivially has compact support (pre-image of Range$\backslash\{0\}$). Therefore, by the Riesz-Markov-Kakutani Representation Theorem, the continuous dual space of $C(Z)$ is the space of regular Borel measures−let us denote it $B(Z)$. Linear functionals on $C(Z)$ have the form: given $v \in B(Z)$

$$\forall f \in C(Z), \quad v(f) = \int_Z f(x)\, dv(x).$$

The norm of $v \in B(Z)$ induced by $C(Z)$ is known as the total variation of v:

$$TV(v) := \sup\left\{ \left| \int_Z g(x)\, dv(x) \right| : \|g\|_Z \le 1, g \in C(Z) \right\}.$$

Going back to the Hilbert space $\mathcal{H}$, given a measure $v \in B(Z)$, we would now like to identify the integral $\int_Z \Phi(x) dv(x)$ as an element of $\mathcal{H}$. To do so, we must call upon our good friend Riesz Representation Theorem. Let us define the (conjugate) linear functional: for each $h \in \mathcal{H}$,

$$L[h] := \int_Z \langle \Phi(x), h \rangle\, dv(x);$$

notice that L is bounded by the Hölder's Inequality and the norms we defined on $C(Z)$ and $B(Z)$:

$$\|L\| \le \|\Phi\|_\infty TV(v) < \infty.$$

Since L is a bounded linear functional, Riesz Representation Theorem tells us that in fact L is *uniquely* determined by an element $w \in \mathcal{H}$ such that $L[h]$ can be re-written as

$$L[h] = \langle w, h \rangle_{\mathcal{H}}.$$

Notice that the only candidate for w is precisely the element $\int_Z \Phi(x)\, dv(x)$:

$$L[h] = \left\langle \int_Z \Phi(x)\, dv(x), h \right\rangle_{\mathcal{H}} = \int_Z \langle \Phi(x), h \rangle_{\mathcal{H}}\, dv(x). \tag{1}$$

Note that the element $\int_Z \Phi(x)\, dv(x)$ depends on choice of $v \in B(Z)$. Let us define a map $U : B(Z) \to \mathcal{H}$

$$U(v) = \int_Z \Phi(x)\, dv(x). \tag{2}$$

Plugging this back into 1 we get

$$\langle U(v), h \rangle_{\mathcal{H}} = \int_Z \langle \Phi(x), h \rangle_{\mathcal{H}}\, dv(x).$$

Now if we set $h = \Phi(y)$, where $y \in Z$, we get the interesting result

$$\langle U(v), \Phi(y) \rangle = \int_Z \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}\, dv(x) := \int_Z k(x, y)\, dv(x).$$

This allows us to bound the norm of the operator U:

$$\begin{aligned} \|U(v)\|_{\mathcal{H}}^2 &= \int_Z \int_Z k(x, y) d\overline{v}(y) dv(x) \\ \implies \|U\|^2 &\leq \|K\|_\infty \\ \|U\| &\leq \sqrt{\|K\|_\infty}. \end{aligned}$$

Since U is a linear operator that is bounded, it is also continuous. Nice!

We are now ready to prove a series of results that relate $K(Z)$ and $C(Z)$ to fundamental subspaces induced by the operator U. Let us denote the nullspace of U:

$$Null(U) := \{v \in B(Z) : U(v) = 0\}.$$

We also introduce the notion of an *annihilator* of $S \subseteq C(Z)$. The annihilator contains all linear functionals from the dual space $B(Z)$ that are uniformly 0 when evaluated on S:

$$Ann_{C(Z)}(S) := \{v \in B(Z) : \int_Z f(x)\, dv(x) = 0 \quad \forall f \in S\}.$$

Observation 2.2 *Since linear functionals are, well, linear, we observe that*

$$Ann_{C(Z)}(S) = Ann_{C(Z)}(\overline{span}(S)).$$

This further implies that two sets have the same annihilator if they belong to the same subspace of $C(Z)$.

Lemma 2.3 *Suppose that Z is a compact set. Then*

$$Ann_{C(Z)}(K(Z)) = Null(U). \tag{3}$$

Corollary 2.4 *$K(Z) = C(Z)$ if and only if U is injective.*

We introduce the following subspace of $\mathcal{H}$:

$$\Phi(Z) := \overline{span}\{\Phi(x) : x \in Z\} = \overline{span}\{k(x, \cdot) : x \in Z\} \subset \mathcal{H},$$

Lemma 2.5 $\overline{R(U)} = \Phi(Z)$.

Provided an orthonormal basis $\Gamma = \{\gamma_1, \gamma_2, \dots, \gamma_i, \dots\}$ for $\mathcal{H}$, we introduce a new subspace of $C(Z)$. Define the functions $f_{\gamma_i} \in C(Z)$ where for $x \in Z$, $f_{\gamma_i}(x) = \langle \Phi(x), \gamma_i \rangle_{\mathcal{H}}$ (essentially γ_i but restricted to Z). We then get a natural subspace of $C(Z)$: [1]

$$\Psi(\Gamma) := \overline{span}\{f_{\gamma_i} : \gamma_i \in \Gamma\}.$$

We are now ready to state a core theorem of the paper.

Theorem 2.6 *Suppose Z is a compact subset of the input space $\mathcal{X}$, k is a kernel with corresponding feature map Φ, and Γ is an orthonormal basis for the RKHS $\mathcal{H}$. Then*

$$K(Z) = \Psi(\Gamma).$$

Since $\Psi(\Gamma)$ is simply the span of the orthonormal basis functions restricted to Z such that they belong to $C(Z)$, we observe that the above theorem essentially tells us that **a kernel is universal if and only if its features are universal**.

This is one way to characterize the universality of a given kernel. We can also look at it from a more spectral perspective. Recall the corresponding Hilbert-Schmidt operator given a positive-definite kernel symmetric kernel:

$$T[f] = \int_Z k(\cdot, y) g(y)\, d\mu(y).$$

Lemma 2.7 *Suppose Z is a compact set, k is a PSD kernel, and $\{\lambda_i\}_{i \in \mathbb{N}} \subset \mathbb{R}^+$ and $\Gamma = \{\phi_i\}_{i \in \mathbb{N}} \subset L^2(Z, \mu)$ are the (positive) eigenvalues and orthonormal eigenvectors corresponding to the diagonalization of the compact symmetric operator T such that Γ is a basis for $\mathcal{H}$. If further* $\text{supp}(\mu) = Z$, *then $K(Z) = C(Z)$ if and only if $\Psi(\Gamma) = C(Z)$* [2].

Since the eigenvectors form an orthonormal basis for $\mathcal{H}$, this lemma is simply an application of the previous theorem. However, it does lead us to the following alternate characterization of universal kernels:

Theorem 2.8 *If* $\text{supp}(\mu) = Z$ *in the Hilbert-Schmidt operator we defined earlier, then $K(Z) = \overline{R(T)}$.*

[1] Honestly the paper's original notation for the following expression is pretty terrible and I'm pretty sure it leads to an error that I will point out later.

[2] The paper originally read $\overline{span}\{\phi_i\}_{i \in \mathbb{N}} = C(Z)$, which seems wrong because the span of the eigenvectors should be the whole space $\mathcal{H}$ by the Spectral Theorem (the whole point of defining the operator T).

Proof: We must simply show $Ann_{C(Z)}K(Z) = Ann_{C(Z)}R(T)$. If $v \in Ann[K(Z)]$, then for each $y \in Z$

$$\int_Z k(x,y)\, dv(x) = 0.$$

Since everything is continuous, we can use Fubini's theorem to re-write

$$\begin{aligned}\int_Z (Tg)(x)\, dv(x) &= \int_Z \left(\int_Z k(x,y) g(y)\, d\mu(y) \right) dv(x) \\ &= \int_Z g(y) \left(\int_Z k(x,y) dv(x) \right) d\mu(y) \\ &= \int_Z g(y) \cdot 0\, d\mu(y) \\ &= 0.\end{aligned}$$

This implies $v \in Ann[R(T)]$.

To prove the other direction, let $v \in Ann[R(T)]$. By the set of equations above, we have that for any $g \in C(Z)$

$$\int_Z (Tg)(x)\, dv(x) = \int_Z g(y) \left(\int_Z k(x,y) dv(x) \right) d\mu(y) = 0.$$

Now we set $g = \overline{\int_Z k(x,\cdot)\, dv(x)}$ and see that this implies

$$\int_Z g(y) \left(\int_Z k(x,y) dv(x) \right) d\mu(y) = \int_Z |g(y)|^2\, d\mu(y) = 0.$$

However, since $\text{supp}(\mu) = Z$, the integral is 0 if and only if $g = 0$, implying that $v \in Ann[K(Z)]$. ∎

We observe that this gives us another characterization of universal kernels that looks at the range of the corresponding Hilbert-Schmidt operator. In other words, if $\overline{R(T)} = C(Z)$, then we know $K(Z) = C(Z)$ and that our kernel is universal. I'm not sure what the immediate benefits of this representation are, but I have found work [4] that discusses reconstructing a continuous integral operator from finite-dimensional data, i.e. matrices. May be unrelated to universal kernels, but I guess I'll give it a look later.

2.2 Notable Examples

The paper starts with an example that is rather close to heart: the dot product kernel. Given an entire (holomorphic on the whole complex plane):

$$G(z) = \sum_{i=0}^{\infty} c_n z^n, \quad c_n > 0$$

we define the kernel

$$k(x,y) = G(\langle x,y\rangle) = \sum_{i=0}^{\infty} c_n \langle x,y\rangle^n, \quad x,y \in \mathbb{C}^n.$$

It is not difficult to verify that the dot product kernel is a reproducing kernel, and that it is universal on $\mathbb{C}^n$ and hence $\mathbb{R}^n$. If we define our feature maps to be polynomial factors, complex analysis (Runge's Theorem and/or Stone-Weierstrass) tells us polynomials well-approximate continuous functions (no poles). With similar reasoning, we can establish that if the function $G(z)$ was instead analytic on the unit disk $\mathcal{D} = \{z \in C : |z| \leq 1\}$, and its power series expansion only contains positive coefficients, then the dot product kernel is universal on the unit ball in n-dimensions.

Another example of a kernel that I was (still am) not familiar with is the "Schoenberg kernel", defined on the unit sphere $S^n = \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$. We consider the k-th degree Gegenbauer polynomials: $P_k^n, k \in \mathbb{Z}^+$, determined by the generating function

$$\frac{1}{(1-2zt+z^2)^{(n-1)/2}} = \sum_{k=1}^{\infty} P_k^d(t) z^k, \quad z \in D,\ t \in [-1,1].$$

Suppose that we have a sequence $\{a_k\}$ such that following sequence converges:

$$\sum_{k=1}^{\infty} a_k P_k^d(1) < \infty.$$

This guarantees that the following function converges uniformly on $[0,\pi]$:

$$g(t) := \sum_{k=1}^{\infty} a_k P_k^d(\cos(t)).$$

Recall that the (geodesic) distance between two points $x, y \in S^n$ is

$$\Delta(x,y) = \arccos(\langle x,y\rangle).$$

The claim is that a jointly continuous function $k : S^n \times S^n \to \mathbb{R}$ is a kernel on S^n if and only if it has the form:

$$k(x,y) := g(\Delta(\langle x,y\rangle)), \quad x,y \in \mathbb{R}^n.$$

Furthermore, k is a universal kernel on S^n if and only if all $a_k > 0$ in the definition of $g(t)$ above.

2.3 Translation Invariant Kernels

We conclude this final project with a discussion of translation invariant kernels, i.e. kernels that have the form

$$k(x,y) = K(x-y), \quad x,y \in \mathbb{R}^n.$$

where K is a continuous function on $\mathbb{R}^n$. Notably, the Gaussian (radial) kernel is a translation invariant kernel. Bochner proved that k is a kernel if and only if there is a unique finite Borel measure μ on $\mathbb{R}^n$ such that K has the form

$$K(x) = \int_{\mathbb{R}^n} e^{i\langle x,y\rangle}\, d\mu(y) \quad \forall x \in \mathbb{R}^n.$$

We observe that the measure μ determines what the kernel will look like. Naturally, we then study the correspondence between the chosen measure and the universality of the resulting kernel. Borrowing the notation established in the earlier parts, we set $\mathcal{X} = \mathbb{R}^n$ and define the Hilbert space $\mathcal{H}$ of all complex-valued functions defined on $\text{supp}(\mu)$ with the corresponding inner product

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\text{supp}(\mu)} f(x)\overline{g(x)}d\mu(x).$$

The natural feature map $\Phi : \mathbb{R}^n \to \mathcal{H}$ is defined

$$\Phi(x) := e^{i\langle x,\cdot\rangle}, \quad \Phi(x)(y) = e^{i\langle x,y\rangle}.$$

We then define the set of exponential features

$$E(\mu) := \{\Phi(x) : x \in \text{supp}(\mu)\}.$$

Just like for an arbitrary kernel, $E(\mu)$ is said to be universal if $E(\mu)$ dense in $C(Z)$ for any compact $Z \subset \mathbb{R}^n$. We can ship in all the universality conditions established for the arbitrary kernel. However, we get some additional properties that I will list here (their proofs follow from some measure theory):

Proposition 2.9 *Given Borel measure μ,*

1. *If* $\text{supp}(\mu)$ *has positive Lebesgue measure on $\mathbb{R}^n$ then the translation kernel k is universal.*

2. *If the continuous part of μ in its Lebesgue decomposition is non-zero then the translation kernel k is universal.*

These two items will help us prove the last, important result of the paper: for basically any choice of measure, the Gaussian kernel is universal.

2.4 The Gaussian (Radial) Kernel

We define a special case of the Schoenberg kernels on $\mathbb{R}^n \times \mathbb{R}^n$ where

$$k(x, y) := g(\|x - y\|^2), \quad x, y \in \mathbb{R}^n.$$

k is a kernel on $\mathbb{R}^n \times \mathbb{R}^n$ for all $n \in \mathbb{N}$ if and only if there exists a finite Borel measure μ on $\mathbb{R}^+$ such that

$$g(t) := \int_{\mathbb{R}^+} e^{-tx} d\mu(x). \tag{4}$$

Theorem 2.10 *If the measure μ in 4 is not concentrated at 0, then the radial kernel k is universal.*

The proof seems pretty involved, and I'll probably work through it eventually out of interest, but as a result of the theorem, we have the following result:

Corollary 2.11 *The following kernels are universal: $\alpha, \beta > 0$*

1. $k(x, y) := e^{-\alpha\|x-y\|^2}, \quad x, y \in \mathbb{R}^n$
2. $k(x, y) := (\beta + \|x - y\|^2)^{-\alpha}, \quad x, y \in R^n.$

The first kernel is our favorite Gaussian kernel.

3 Final Thoughts

This paper introduced a lot of core machinery to prove the good properties of many important kernels, which has involved a lot of measure theory and functional analysis (cool stuff). Some natural follow-up reading would be papers that actually need numerical bounds on how well universal kernels perform, as it's clear some have approximations that converge far faster than others. Also, as mentioned in the paper, in practice there is some nuance in approximating a function by the kernel or by its features, which is not something intuitively clear to me right now. I also plan to write up some stuff on kernel ridge regression with Sobolev kernels in my own time, which I will append here at some point.

References

[1] Nachman Aronszajn. Theory of reproducing kernels. Transactions of the American Mathematical Society 68:337-404, 1950.

[2] Arthur Gretton. Introduction to RKHS, and some simple kernel algorithms, 2017.

[3] C. A. Micchelli, Y. Xu, H. Zhang. Universal Kernels. Journal of Machine Learning Research 7:2651-2667, 2006.

[4] L. Rosasco, M. Belkin, E. De Vito. On Learning with Integral Operators. Journal of Machine Learning Research 11:905-934, 2010.