MERCER THEORY: A SURVEY

RAHUL KRISHNA

1. Introduction

In this expository note, we present two forms of Mercer's theorem, and discuss their relevance to the "feature map" interpretation of the kernel trick.

Using this as a jumping off point, we then explain how the kernel functions appearing in Mercer theory can give rise to a different, easier, construction of a feature map. This leads us to the notion of a *reproducing kernel Hilbert space (RKHS)*, and the very general - and somewhat formal - theorem running this machine: the Moore-Aronszajn theorem.

We have tried to make a point of keeping this note relatively self-contained. We assume that the reader is familiar with measure theory and point-set topology, as well as the basic definitions occurring in the study of function spaces - Hilbert spaces, Banach spaces, bounded operators, etc. We assume too some level of comfort with "abstract" linear algebra (in particular, use of the tensor product) as we find such language extremely clarifying. Beyond this, we have tried to provide most of the background.

Throughout, all functions are real-valued. Unless explicitly stated, all Hilbert spaces are real. *We do not assume that our Hilbert spaces are separable.*[1]

2. A brief reminder on (non-separable) Hilbert spaces

Definition 2.1. Let $H = (H, (\cdot,\cdot))$ be a Hilbert space. A set of vectors $\{e_\alpha\}_{\alpha\in I}$ is *orthonormal* if

$$(e_\alpha, e_\beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{if } \alpha \neq \beta \end{cases}.$$

Note that in the above, I can be uncountably infinite.

A maximal orthonormal set in H is called an *orthonormal basis* of H.

By Zorn's lemma, every Hilbert space H admits an orthonormal basis.

Lemma 2.2. *(Bessel's inequality) Let $\{e_1, \dots, e_n\}$ be a finite orthonormal set in a Hilbert space H. Then, for every $\phi \in H$*

$$\sum_{i=1}^{n} (\phi, e_i)^2 \leq \|\phi\|^2.$$

Proof.

$$0 \leq (\phi - \sum_{i=1}^{n}(\phi, e_i)e_i, \phi - \sum_{i=1}^{n}(\phi, e_i)e_i) = (\phi,\phi) - \sum_{i=1}^{n}(\phi, e_i)^2.$$

□

Corollary 2.3. *Let H be a Hilbert space, and let $\{e_\alpha\}_{\alpha\in I}$ be an orthonormal basis of H. Then, for any $\phi \in H$,*

$$\sum_{\alpha\in I} (\phi, e_\alpha)^2 \leq \|\phi\|^2 < \infty.$$

In particular, all but countably many of the (ϕ, e_α) are zero.

---

[1] Non-separable Hilbert spaces can appear in "real life." E.g. if $\mathcal{X} = \mathrm{U}(1)^{\mathbb{R}}$ with the product topology, then $\mathcal{X}$ is a compact Abelian group, so admits a Haar probability measure. $H = L^2(\mathcal{X})$ is not separable - there is no countable family of opens shrinking to 1. $\mathcal{X}$ is also not first countable.

Proof. For any finite set $S \subseteq I$, we have by Lemma 2.2

$$\sum_{\alpha \in S} (\phi, e_\alpha)^2 \leq \|\phi\|^2 < \infty.$$

Taking the supremum over all finite $S \subseteq I$, we get

$$\sum_{\alpha \in I} (\phi, e_\alpha)^2 = \sup_{\substack{S \subseteq I \\ \#S < \infty}} \sum_{\alpha \in S} (\phi, e_\alpha)^2 \leq \|\phi\|^2 < \infty.$$

□

Proposition 2.4. *(Hilbert's projection theorem) Let H be a Hilbert space, $H' \subseteq H$ a closed subspace. Given $\phi \in H$, let $d(\phi, H') := \inf_{\psi \in H'} \|\phi - \psi\|$.*

(1) There is a unique $\phi' \in H'$ so that $\|\phi - \phi'\| = d(\phi, H')$.
(2) $(\phi - \phi', \psi) = 0$ for every $\psi \in H'$.

This unique ϕ' is called *the orthogonal projection* of ϕ onto H'; we write $\phi' = \text{proj}_{H'}\phi$.

Proof. (1) Uniqueness is easy to establish. Suppose that

$$\|\phi - \phi_1'\| = d(\phi, H') = \|\phi - \phi_2'\|$$

Consider the quadratic function of a single real variable t

$$q(t) := \|\phi - t\phi_1' - (1-t)\phi_2'\|^2 = \|\phi_2' - \phi_1'\|^2 t^2 + \text{ lower degree terms}$$

By assumption, this achieves its global minimum $d(\phi, H')^2$ at $t = 0$ and $t = 1$ - it must therefore be constant. In particular, the coefficient of t^2 vanishes, i.e. $\phi_1' = \phi_2'$.

As for existence of such a closest vector ϕ', let $\psi_n \in H'$ be a sequence such that

$$\lim_{n \to \infty} \|\phi - \psi_n\| = d(\phi, H').$$

Note that for any n, m,

$$\|\psi_n - \psi_m\|^2 + 4\|\phi - \frac{\psi_n + \psi_m}{2}\|^2 = 2\|\phi - \psi_n\|^2 + 2\|\phi - \psi_m\|^2.$$

Given any $\varepsilon > 0$, there exists an N so that if $n, m \geq N$,

$$2\|\phi - \psi_n\|^2 + 2\|\phi - \psi_m\|^2 - 4d(\phi, H')^2 < \varepsilon^2.$$

Thus for $n, m \geq N$,

$$\|\psi_n - \psi_m\|^2 \leq \|\psi_n - \psi_m\|^2 + 4\|\phi - \frac{\psi_n + \psi_m}{2}\|^2 - 4d(\phi, H')^2 < \varepsilon^2$$

since $(\psi_n + \psi_m)/2 \in H'$, and so

$$\|\phi - \frac{\psi_n + \psi_m}{2}\| \geq d(\phi, H'.)$$

It follows that $\{\psi_n\}_{n=1}^\infty$ is Cauchy. Taking $\phi' = \lim_{n \to \infty} \psi_n$ suffices.

(2) Let ψ' be any vector in H', and let $\alpha = (\phi - \phi', \psi')$. For $t \in \mathbb{R}$, consider

$$\|\phi - \phi' - t\psi'\|^2 = (\phi - \phi' - t\psi', \phi - \phi' - t\psi') = d(\phi, H')^2 - 2\alpha t + t^2\|\psi'\|^2.$$

If $\alpha \neq 0$, then there exists t so that $t^2\|\psi'\|^2 - 2\alpha t < 0$, which would violate (1). Thus, $\alpha = 0$. □

Remark 2.5. If $\phi' \in H'$ is such that $(\phi - \phi', \psi) = 0$ for every $\psi \in H'$ (i.e. satisfies (2)), then $\phi' = \text{proj}_{H'}(\phi)$. This is because, for every $\psi \in H'$.

$$\|\phi - \psi\|^2 = \|(\phi - \phi') + (\phi' - \psi)\|^2 = \|\phi - \phi'\|^2 + \|\phi' - \psi\|^2$$

by the Pythagorean theorem, and this quantity is greater than $\|\phi - \phi'\|^2$ unless $\psi = \phi'$.

Remark 2.6. (1) and its proof above apply equally well to the more general situation of a closed convex subset in H (i.e. not necessarily a linear subspace).

Lemma 2.7. *Let H be a Hilbert space, $V \subseteq H$ a subspace. Let $V^\perp = \{\phi \in H : (\psi, \phi) = 0$ for all $\psi \in V\}$. Then $V^\perp$ is closed, and $(V^\perp)^\perp$ is the closure of V in H.*

Proof. For a given ψ, $(\mathbb{R}\psi)^\perp$ is visibly closed, as it is the preimage of the closed set $\{0\}$ under the continuous linear functional $\phi \mapsto (\psi, \phi)$. Thus,

$$V^\perp = \bigcap_{\psi \in V} (\mathbb{R}\psi)^\perp$$

must be closed. Since it is apparent that $V \subseteq (V^\perp)^\perp$, it follows that the closure $\overline{V} \subseteq (V^\perp)^\perp$. It remains to see the reverse inclusion. By replacing V with its closure, we may assume V is closed. Let $\phi \in (V^\perp)^\perp$, and let ψ be its orthogonal projection onto V. Then $\phi - \psi \in V^\perp$, so

$$(\phi - \psi, \phi - \psi) = (\phi, \phi - \psi) - (\psi, \phi - \psi) = 0 - 0 = 0,$$

i.e. $\phi = \psi$. □

Corollary 2.8. *Let H be a Hilbert space, $H' \subseteq H$ a closed subspace. Then given $\phi \in H$, there are unique $\phi' \in H'$ and $\phi'' \in H'^\perp$ so that $\phi = \phi' + \phi''$.*

Proof. Uniqueness is apparent, since $H' \cap H'^\perp = \{0\}$ (by positive definiteness of the inner product). Let ϕ' be the orthogonal projection of ϕ onto H'. Then $\phi'' := \phi - \phi' \in H'^\perp$ by (2) of Hilbert's projection theorem. □

Lemma 2.9. *Let H be a Hilbert space, $\{e_\alpha\}_{\alpha \in I}$ an orthonormal basis. Then the algebraic span of the e_α (i.e. the space of finite linear combinations of the e_α) is dense in H.*

Proof. Let H' be the closure in H of the algebraic span of the e_α. Let $e \in H'^\perp$. $(e, e_\alpha) = 0$ for all $\alpha \in I$. By maximality of $\{e_\alpha\}_{\alpha \in I}$, $e = 0$. □

Corollary 2.10. *Let H be a Hilbert space, and let $\{e_\alpha\}_{\alpha \in I}$ be an orthonormal basis of H. Then*

$$\phi = \sum_{\alpha \in I} (\phi, e_\alpha) e_\alpha.$$

(All but countably many terms in this sum are 0.)

Proof. Let $J = \{\beta \in I : (\phi, e_\beta) \neq 0\}$. J is at most countable. Let H' be the closure of the algebraic span of $\{e_\beta\}_{\beta \in J}$, $H'' = (H')^\perp$ its orthogonal complement. By Corollary 2.8 ϕ can be uniquely written as $\phi = \phi' + \phi''$, with $\phi' \in H'$, $\phi'' \in H''$.

We claim that $\phi'' = 0$. For all $\beta \in J$, $(\phi'', e_\beta) = 0$; if $\alpha \in I - J$, $(\phi'', e_\alpha) = (\phi, e_\alpha) - (\phi', e_\alpha) = 0$ as well. Thus $\phi'' = 0$ by Lemma 2.9.

It remains to check that $\phi = \phi'$ has the desired form. This is easy: consider the expression

$$\sum_{\beta \in J} (\phi, e_\beta) e_\beta.$$

By Bessel's inequality, this defines an element of H' - if we look at

$$\phi - \sum_{\beta \in J} (\phi, e_\beta) e_\beta$$

then by the same logic as above, this difference is orthogonal to all e_α, hence must be 0. □

Proposition 2.11. *(The Riesz representation theorem) Let H be a Hilbert space.*

(1) If $\psi \in H$, the linear functional

$$(\psi, \cdot) : \phi \mapsto (\psi, \phi)$$

is bounded, with operator norm equal to $\|\psi\|$.

(2) If l is a bounded linear functional on H, then there is a unique $\psi \in H$ so that $l = (\psi, \cdot)$.

Proof. (1) is apparent: Cauchy-Schwartz gives

$$|(\psi, \phi)| \leq \|\psi\| \|\phi\|$$

which shows that the operator norm of $(\psi, \cdot)$ is bounded by $\|\psi\|$; equality follows from taking $\phi = \psi$.

As for (2), if $l = 0$ there is nothing to prove. If $l \neq 0$, let ϕ_0 be such that $l(\phi_0) \neq 0$. Let $H' = \ker(l)$; by the continuity of l, H' is a closed subspace. Let ϕ_0' ben the orthogonal projection of ϕ_0 to H', and $\phi_0'' = \phi_0 - \phi_0' \in (H')^\perp$. Note that $\phi_0'' \neq 0$.

We claim that ϕ_0'' spans $(H')^\perp$. Let $\varphi'' \in (H')^\perp$, and let $c \in \mathbb{R}$ be such that $l(\phi'') = cl(\phi_0'')$. Then $\phi'' - c\phi_0'' \in \ker(l) = H'$, hence is 0.

So now, finally, let $a \in \mathbb{R}^\times$ be such that $l(\phi_0'') = l(\phi_0) = a(\phi_0, \phi_0) = a(\phi_0, \phi_0'') = a(\phi_0'', \phi_0'')$, and call $\psi = a\phi_0''$. Let $\phi \in H$ be any vector, and write $\phi = \phi' + b\phi_0''$ for $b \in \mathbb{R}$. Then

$$(\psi, \phi) = (a\phi_0'', b\phi_0'') = ab(\phi_0'', \phi_0'')$$

while

$$l(\phi) = bl(\phi_0'') = ab(\phi_0'', \phi_0'').$$

□

Proposition 2.12. *Let $T : H \to H$ be a bounded linear transformation. Then there is a unique bounded linear transformation $T^* : H \to H$ so that*

$$((\psi, T\phi) = (T^*\psi, \phi)$$

for any $\psi, \phi \in H$. Moreover, $\|T\| = \|T^*\|$.

Proof. We appeal to Proposition 2.11 to define T^*, namely for each $\psi \in H$, the linear functional

$$\phi \mapsto (\psi, T\phi)$$

is represented by a vector we call $T^*\psi$. We have

$$(T^*(c_1\psi_1 + c_2\psi_2), \phi) = (c_1\psi_1 + c_2\psi_2, T\phi) = c_1(\psi_1, T\phi) + c_2(\psi_2, T\phi) = (c_1T^*\psi_1 + c_2T^*\psi_2, \phi).$$

Since this holds for all ϕ, we have $(T^*(c_1\psi_1 + c_2\psi_2) = c_1T^*\psi_1 + c_2T^*\psi_2$ so T^* is linear. To see that T^* is bounded, we compute

$$|(T^*\psi, \phi)| = |(\psi, T\phi)| \leq \|\psi\|\|T\|\|\phi\|$$

Plugging in $\phi = T^*\psi$, we get

$$\|T^*\psi\| \leq \|T\|\|\psi\|.$$

so that $\|T^*\| \leq \|T\|$. Applying this same identity with T replaced with T^*, using that $T^{**} = T$, we obtain the reverse inequality. □

Definition 2.13. We call T^* the *adjoint* of T.

3. Mercer's theorem: measurable form

Let $\mathcal{X} = (\mathcal{X}, \mathfrak{M}, \mu)$ be a σ-finite measure space. The space $L^2(\mathcal{X})$ is a (possibly non-separable) Hilbert space. We use $\|\phi\| = (\phi, \phi)^{1/2}$ to denote the norm on $L^2(\mathcal{X})$.

3.1. **Hilbert-Schmidt kernels and operators.**

3.1.1. *Square integrable kernels.* Let K be a measurable function on $\mathcal{X} \times \mathcal{X}$, square integrable on $\mathcal{X} \times \mathcal{X}$. Such a K is often called an L^2 *kernel* on $\mathcal{X}$. To K, we associate, for each measurable and square integrable ϕ on $\mathcal{X}$, a function

$$T_K\phi(x) = \int_{\mathcal{X}} K(x, y)\phi(y)dy.$$

Note that, by Cauchy-Schwartz,

$$|T_K\phi(x)| = \left|\int_{\mathcal{X}} K(x, y)\phi(y)dy\right| \leq \|K(x, \cdot)\|\|\phi\| \tag{3.1}$$

and since

$$\int_{\mathcal{X}} \|K(x, \cdot)\|^2 dx = \|K\|^2_{L^2(\mathcal{X}\times\mathcal{X})} < \infty,$$

$\|K(x, \cdot)\|$ is finite for almost every $x \in \mathcal{X}$. It follows that $T_K\phi(x)$ is finite for almost every $x \in \mathcal{X}$, hence well defined a.e.. It is also visibly measurable, as it is defined as the integral of a measurable function.

Lemma 3.1. *We have the following:*

(1) *$T_K\phi$ is square integrable on $\mathcal{X}$, and in fact*

$$\|T_K\phi\| \leq \|K\|_{L^2(\mathcal{X}\times\mathcal{X})}\|\phi\|.$$

(2) If ϕ_1, ϕ_2 are square integrable measurable functions on $\mathcal{X}$ which agree a.e., then $T_K\phi_1$ and $T_K\phi_2$ also agree a.e.. Thus, T_K descends to a linear map, called the Hilbert-Schmidt operator associated with K

$$T_K : L^2(\mathcal{X}) \to L^2(\mathcal{X}).$$

(3) Let K_1, K_2 be measurable and square integrable on $\mathcal{X} \times \mathcal{X}$, with corresponding operators T_{K_1} and T_{K_2}. If K_1, K_2 agree a.e. on $\mathcal{X} \times \mathcal{X}$, then $T_{K_1}\phi$ and $T_{K_2}\phi$ agree a.e. on $\mathcal{X}$, for every $\phi \in L^2(\mathcal{X})$.

Proof. (1) This follows immediately from (3.1):

$$\int_{\mathcal{X}} |T_K\phi(x)|^2 dx \leq \int_{\mathcal{X}} \|K(x,\cdot)\|^2 \|\phi\|^2 dx = \|K\|^2_{L^2(\mathcal{X}\times\mathcal{X})} \|\phi\|^2.$$

(2) It is enough to show, by considering $\phi = \phi_1 - \phi_2$, that if a function ϕ is zero a.e., then $T_K\phi = 0$. But this is apparent: the integral of an a.e. 0 function is 0. (3) Call $K = K_1 - K_2$. K is zero almost everywhere on $\mathcal{X} \times \mathcal{X}$, and hence

$$\int_{\mathcal{X}\times\mathcal{X}} |K(x,y)| dx dy = 0.$$

By Fubini's theorem, for almost every x, $K(x,\cdot)$ is measurable, and in fact is 0 a.e. (in y). Thus, for a.e. x, $T_K\phi(x) = 0$. □

Remark 3.2. For K a square integrable kernel, let

$$K^*(x,y) := K(y,x).$$

Then

$$T_K^* = T_{K^*}.$$

For any Hilbert space H, we let $\mathrm{End}(H)$ be the vector space of bounded endomorphisms of H. This is a Banach space, with norm given by the operator norm:

$$\|T\|_{\mathrm{End}(H)} := \sup_{\|\phi\|=1} \|T\phi\|.$$

When there is no risk of confusion, we simply write $\|T\|$ for $\|T\|_{\mathrm{End}(H)}$.

Consider the linear map (of Banach spaces)

$$T_\bullet : L^2(\mathcal{X} \times \mathcal{X}) \to \mathrm{End}(L^2(\mathcal{X}))$$
$$K \mapsto T_K.$$

By Lemma 3.1, $T_\bullet$ is well-defined.

Lemma 3.3. *$T_\bullet$ has the following basic properties:*

(1) We have

$$\|T_K\| \leq \|K\|_{L^2(\mathcal{X}\times\mathcal{X})}.$$

In particular, $T_\bullet$ is bounded (i.e. continuous).

(2) If $L^2(\mathcal{X})$ is separable, then $T_\bullet$ is also injective.

Proof. (1) is immediate from Lemma 3.1. (2) Pick a countable orthonormal basis $\{e_n\}_n$ of $L^2(\mathcal{X})$. Let $K \in L^2(\mathcal{X} \times \mathcal{X})$ satisfy $T_K = 0$. Then for each n, there is a measure 0 set E_n so that

$$T_K e_n(x) = 0$$

for all $x \notin E_n$. Let $E = \bigcup_n E_n$; note that $\mu(E) = 0$. Then for all x outside of E, $K(x,\cdot)$ is orthogonal to e_n, hence for all such x, $K(x,\cdot)$ is 0 in $L^2(\mathcal{X})$. It follows that $\|K\|_{L^2(\mathcal{X}\times\mathcal{X})} = 0$. □

Remark 3.4. We will see in a moment that unless $L^2(\mathcal{X})$ is finite dimensional, $T_\bullet$ is never surjective.

Definition 3.5. Regardless of whether $L^2(\mathcal{X})$ is separable or not, we call the image of $T_\bullet$ the space of *Hilbert-Schmidt operators* on $L^2(\mathcal{X})$, and denote it by $\mathrm{HS}(L^2(\mathcal{X}))$.

3.1.2. *Hilbert-Schmidt operators on a Hilbert space H.* Let us make a key conceptual definition that will help to abstract away a few confusing points.

Definition 3.6. Let H, H' be Hilbert spaces. Let $H \otimes H'$ be the usual algebraic tensor product, i.e. the space consisting of finite linear combinations of pure tensors $\phi \otimes \phi'$, $\phi \in H, \phi' \in H'$. Then $H \otimes H'$ is a pre-Hilbert space, with inner product given by setting

$$(\phi \otimes \phi', \psi \otimes \psi')_{H \otimes H'} = (\phi, \psi)_H (\phi', \psi')'_H$$

and extending by (bi-)linearity. We set $H \widehat{\otimes} H' := (\widehat{H \otimes H'})$ to be the Hilbert space completion of the algebraic tensor product, and call it the *completed tensor product of H and H'.*

Remark 3.7. If $\{e_\alpha\}_{\alpha \in I}$ and $\{e'_\beta\}_{\beta \in I'}$ are orthonormal bases of H and H' respectively, then $\{e_\alpha \otimes e'_\beta\}$ is an orthonormal basis of $H \widehat{\otimes} H'$.

We now examine the relationship between $H \widehat{\otimes} H$ and $\mathrm{End}(H)$.

First, the familiar finite dimensional situation. Let V be a finite dimensional inner product space. The inner product identifies V and its dual $V^\vee$. We also have the canonical isomorphism $a : V^\vee \otimes V \to \mathrm{End}(V)$ (a is for "action"), defined on pure tensors by $a(l \otimes \phi) = l(\cdot)\phi$. Together these lead to a composition of canonical isomorphisms

$$V \otimes V \underset{(\cdot,\cdot)}{\xrightarrow{\sim}} V^\vee \otimes V \underset{a}{\xrightarrow{\sim}} \mathrm{End}(V).$$

The story is not as tidy when our Hilbert space H is not required to be finite dimensional. The inner product, by the Riesz representation theorem, isometrically identifies H and the dual Hilbert space $H^\vee = \mathrm{Hom}_{\mathrm{cont}}(H, \mathbb{R})$. Again, there is an injective linear map

$$a : H^\vee \otimes H \to \mathrm{End}(H)$$

defined by sending pure tensors to rank ≤ 1 endomorphisms.

Lemma 3.8. *a is a bounded linear map; more precisely, given an element $\sum_{i=1}^n l_i \otimes \phi_i \in H^\vee \otimes H$,*

$$\|a(\sum_{i=1}^n l_i \otimes \phi_i)\|_{\mathrm{End}(H)} \leq \|\sum_{i=1}^n l_i \otimes \phi_i\|_{H^\vee \otimes H}.$$

Proof. By Gram-Schmidt on the finite set of vectors $\{\phi_i\}_{i=1}^n$, we may assume that $\phi_i = e_i$ are all orthonormal. Given ϕ in H with $\|\phi\| = 1$, we wish to show

$$\|a(\sum_{i=1}^n l_i \otimes e_i)(\phi)\|^2 \leq \|\sum_{i=1}^n l_i \otimes e_i\|^2_{H^\vee \otimes H}.$$

But the left hand side is $\sum_{i=1}^n |l_i(\phi)|^2$, while the right hand side is $\sum_{i=1}^n \|l_i\|^2$, so this inequality obviously holds. □

Since $H^\vee \otimes H$ is dense in $H^\vee \widehat{\otimes} H$, we may extend by continuity to find an injective bounded linear map of Banach spaces $\widehat{a} : H^\vee \widehat{\otimes} H \to \mathrm{End}(H)$. Composing gives

$$H \widehat{\otimes} H \underset{(\cdot,\cdot)}{\xrightarrow{\sim}} H^\vee \widehat{\otimes} H \underset{\widehat{a}}{\to} \mathrm{End}(V).$$

The key point is this: *when H is infinite dimensional, $\widehat{a}$ is far from being surjective.* For instance, the identity operator is Hilbert-Schmidt if and only if H is finite dimensional. We give the image a name.

Definition 3.9. $\mathrm{HS}(H) := \widehat{a}(H^\vee \widehat{\otimes} H)$ is the space of *Hilbert-Schmidt* operators on H.

Remark 3.10. Unless H is finite dimensional, $\mathrm{HS}(H)$ is not a closed space of $\mathrm{End}(H)$ (viewed with the operator norm). Of course, endowed with the *Hilbert-Schmidt norm* $\|\cdot\|_{\mathrm{HS}(H)}$ (the norm it inherits as the image of $H^\vee \widehat{\otimes} H$), it is complete.

3.1.3. *Back to $L^2(\mathcal{X})$.* There is always a map from the algebraic tensor product

$$m : L^2(\mathcal{X}) \otimes L^2(\mathcal{X}) \to L^2(\mathcal{X} \times \mathcal{X});$$

it is defined on pure tensors by sending $\phi \otimes \psi$ to $(x, y) \mapsto \phi(x)\psi(y)$, and extended by linearity to all of $L^2(\mathcal{X}) \otimes L^2(\mathcal{X})$. m is visibly isometric, and so induces an isometric injection with closed image

$$m : L^2(\mathcal{X})\widehat{\otimes} L^2(\mathcal{X}) \hookrightarrow L^2(\mathcal{X} \times \mathcal{X}).$$

Definition 3.11. We call the image of m the space of *Hilbert-Schmidt* kernels on K, and denote it by $L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})$.

Lemma 3.12. *We have*

$$L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})^\perp = \ker T_\bullet.$$

Proof. $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})^\perp$ if and only if, for every pure tensor $\phi \otimes \psi$,

$$(m(\phi \otimes \psi, K))_{L^2(\mathcal{X}\times\mathcal{X})} = 0.$$

In other words, $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})^\perp$ if and only if

$$0 = \int_{\mathcal{X}}\int_{\mathcal{X}} \phi(x)\psi(y)K(x, y)dxdy = (\phi, T_K\psi)$$

for every ϕ and every ψ. But $(\phi, T_K\psi) = 0$ for all ϕ and ψ if and only if $T_K = 0$, as desired. $\square$

Combining Lemma 3.12 and Lemma 3.3, we find

Corollary 3.13. *If $L^2(\mathcal{X})$ is separable, then $L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X}) = L^2(\mathcal{X} \times \mathcal{X})$.*

We summarize our discussion thus far with the following diagram.

$$\begin{tikzcd}
 & L^2(\mathcal{X} \times \mathcal{X}) \arrow[d, "T_\bullet"] & \\
L^2(\mathcal{X})\widehat{\otimes} L^2(\mathcal{X}) \arrow[ur, hook, "m"] \arrow[r, "\sim"] & \mathrm{HS}(L^2(\mathcal{X})) \arrow[r, hook] & \mathrm{End}(L^2(\mathcal{X}))
\end{tikzcd}.$$

When $L^2(\mathcal{X})$ is separable, this diagram simplifies further: all arrows in the left triangle become isomorphisms.

Proposition 3.14. *Let $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})$, and let $\{e_\alpha\}_{\alpha\in I}$ be an orthonormal basis of $L^2(\mathcal{X})$. Then*

$$K(x, y) = \sum_{\alpha\in I} T_K e_\alpha(x)e_\alpha(y)$$

with convergence in the L^2-norm.

Remark 3.15. As usual, all but countably many terms in this sum are non-zero.

Proof. Call $e_{\alpha,\beta} = e_\alpha \otimes e_\beta$; clearly $\{e_{\alpha,\beta}\}_{(\alpha,\beta)\in I\times I}$ is an orthonormal basis of $L^2(\mathcal{X})\widehat{\otimes} L^2(\mathcal{X})$. Since $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X}) = m(L^2(\mathcal{X})\widehat{\otimes} L^2(\mathcal{X}))$, if we call $c_{\alpha,\beta} = (K, m(e_{\alpha,\beta}))$, then we may write

$$K(x, y) = \sum_{\alpha,\beta\in I} c_{\alpha,\beta}e_\alpha(x)e_\beta(y)$$

with convergence in the L^2-norm - in particular, $\sum_{\alpha,\beta} |c_{\alpha,\beta}|^2 < \infty$. But

$$T_K e_\alpha(x) = (K(x, \cdot), e_\alpha) = \sum_{\alpha'\in I} c_{\alpha',\alpha}e_{\alpha'}(x)$$

so

$$\sum_{\alpha\in I} T_K e_\alpha(x)e_\alpha(y) = \sum_{\alpha\in I}\sum_{\alpha'\in I} c_{\alpha',\alpha}e_{\alpha'}(x)e_\alpha(y) = K(x, y).$$

$\square$

3.1.4. *Compactness of Hilbert-Schmidt operators.* Let H be a Hilbert space, $T \in \mathrm{End}(H)$. Let $B = B_{\leq 1}(0)$ be the closed unit ball in H.

Definition 3.16. T is *compact* if $\overline{T(B)}$ is compact. We denote by $\mathrm{End}_{\mathrm{cm}}(H)$ the collection of compact operators on H.

Remark 3.17. Recall from point-set topology that for a metric space X, TFAE:
(1) X compact (every open cover has a finite subcover)
(2) X is sequentially compact (every sequence has a convergent subsequence)
(3) X is complete (every Cauchy sequence converges) and totally bounded (for any $\varepsilon > 0$, there are finitely many ε open balls covering X).

We will often check compactness of a set by checking sequential compactness without further mention.

Remark 3.18. Every finite rank element of $\mathrm{End}(H)$ is compact. If $T \in \mathrm{End}(H)$ is finite rank, then $T(H)$ is a finite dimensional Banach space; as $T(B)$ is contained in a ball in this $\mathbb{R}^n$, its closure is compact.

Lemma 3.19. $\mathrm{End}_{\mathrm{cm}}(H)$ *is a two-sided ideal of* $\mathrm{End}(H)$.

Proof. To see that the sum of two compact operators $T_1 + T_2$ is compact, simply use sequential compactness - if $\{\phi_n\}_n$ is a sequence in B, first pass to a subsequence to ensure that that images under T_1 converge in H, then pass to a further subsequence to ensure that the images under T_2 also converge.

To see the absorbing property, let T be compact, S bounded. Clearly ST is compact, as the continuous image of a compact set is compact. TS is also bounded: the image of the the unit ball under S is contained in the scaled ball $\|S\|B$, and so the image of this under T is must have compact closure. □

Lemma 3.20. *Let H be a Hilbert space, and $T_n \in \mathrm{End}(H)$ a sequence of compact operators converging in the operator norm to a $T \in \mathrm{End}(H)$. Then T is compact.*

Proof. Let $\{\phi_k\}_{k=1}^{\infty}$, be a sequence with $\|\phi_k\| \leq 1$ for all k. We wish to find an infinite subset of indices $S \subseteq \mathbb{N}$ so that

$$\lim_{\substack{k \in S \\ k \to \infty}} T\phi_k$$

converges.

Since each operator T_n is compact, we can find nested infinite subsets of indices $S_1 \supseteq S_2 \supseteq \dots$, $S_n = \{k_1^{(n)}, k_2^{(n)}, \dots\}$ so that

$$\lim_{i \to \infty} T_n \phi_{k_i^{(n)}} \text{ exists.}$$

Let $S = \{k_1, k_2, \dots\}$ be the diagonalized sequence of indices, i.e. k_1 is the first element of S_1, k_2 is the second element of S_2, and so on. We claim that $\{T\phi_{k_j}\}_j$ is a Cauchy sequence.

To see this, let $\varepsilon > 0$. Let N be large enough that $\|T - T_n\| < \varepsilon/3$ for all $n \geq N$. Note that for any k, and $n \geq N$,

$$\|T\phi_k - T_n\phi_k\| \leq \|T - T_n\| < \frac{\varepsilon}{3}.$$

Choose some $n \geq N$. There exists $I = I_n \in S_n$ so that whenever $k_i^{(n)}, k_j^{(n)} \geq I$,

$$\|T_n\phi_{k_i^{(n)}} - T_n\phi_{k_j^{(n)}}\| < \frac{\varepsilon}{3}$$

Let $K \in S$ be such that $K \geq I$. Then for all k_i, k_j in S with $k_i, k_j \geq K$, we have

$$\|T\phi_{k_i} - T\phi_{k_j}\| \leq \|T\phi_{k_i} - T_n\phi_{k_i}\| + \|T_n\phi_{k_i} - T_n\phi_{k_j}\| + \|T_n\phi_{k_j} - T\phi_{k_j}\| < \varepsilon.$$

Thus, the subsequence is Cauchy, hence converges. □

Remark 3.21. It follows that $\mathrm{End}_{\mathrm{cm}}(H) \subseteq \mathrm{End}(H)$ is a closed subspace, i.e. a Banach space in its own right.

Corollary 3.22. *If T is the limit in $\mathrm{End}(H)$, under the operator norm, of a sequence T_n of finite rank operators, then T is compact.*

Proof. Finite rank operators are compact. □

Corollary 3.23. *Every Hilbert-Schmidt operator T is compact.*

Proof. Hilbert-Schmidt operators are by definition the closure of the finite rank operators under the Hilbert-Schmidt norm. But if a sequence converges under the Hilbert-Schmidt norm, it does under the operator norm by Lemma 3.8. □

3.2. **Symmetry and the spectral theorem.** Let H be a Hilbert space, $T \in \mathrm{End}(H)$.

Definition 3.24. T is *symmetric* if

$$(T\phi, \psi) = (\phi, T\psi)$$

for all $\phi, \psi \in H$.

Remark 3.25. Here we only work with bounded (everywhere defined) operators, so the more subtle distinctions between the terms "self-adjoint" and "symmetric" are wholely irrelevant.

Let $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})$ be a Hilbert-Schmidt kernel on $\mathcal{X}$.

Definition 3.26. K is *symmetric* if, for all (x, y) outside of a set of measure 0, $K(x, y) = K(y, x)$.

Remark 3.27. If K is symmetric in sense of Def. 3.26, T_K is symmetric in the sense of Def. 3.24:

$$\begin{aligned}
(T_K\phi, \psi) &= \int_{\mathcal{X}}\int_{\mathcal{X}} K(x, y)\phi(y)dy\psi(x)dx \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} K(y, x)\psi(x)dx\phi(y)dy \\
&= (\phi, T_K\psi).
\end{aligned}$$

(The converse holds as well - it is a consequence of the spectral theorem, so we postpone discussion of it.)

Lemma 3.28. *Let $T \in \mathrm{End}(H)$ be symmetric, and let λ be an eigenvalue of T. Then $\lambda \in \mathbb{R}$.*

Proof. To work with possibly complex eigenvalues and eigenvectors, we must complexify. Complexify H to a complex Hilbert space $H \otimes_{\mathbb{R}} \mathbb{C} = H_{\mathbb{C}}$, with Hermitian inner form $(\cdot, \cdot)_{\mathbb{C}}$, conjugate linear in the first variable, and given by

$$(\phi_1 + i\phi_2, \psi_1 + i\psi_2)_{\mathbb{C}} = ((\phi_1, \psi_1) + (\phi_2, \psi_2)) + i((\phi_1, \psi_2) - (\phi_2, \psi_1)).$$

Since T is symmetric, it naturally extends to a Hermitian bounded operator on $H_{\mathbb{C}}$. Given a complex eigenvalue λ, with non-zero complex eigenvector u, we can compute

$$\overline{\lambda}(u, u) = (T\phi, \phi) = (u, Tu) = \lambda(u, u).$$

It follows that $\overline{\lambda} = \lambda$. □

Lemma 3.29. *Let $T \in \mathrm{End}(H)$ be symmetric. Then*

$$\|T\| = \sup_{\|\phi\|=1} |(T\phi, \phi)|$$

Proof. One inequality is easy: by definition

$$\|T\| = \sup_{\|\phi\|=1} \|T\phi\|$$

so by Cauchy-Schwartz

$$\sup_{\|\phi\|=1} |(T\phi, \phi)| \leq \sup_{\|\phi\|=1} \|T\phi\| = \|T\|.$$

For the other direction, for any ϕ, ψ we have (by symmetry of T)

$$\begin{aligned}
4|(T\phi, \psi)| &= |(T(\phi + \psi), \phi + \psi) - (T(\phi - \psi), \phi - \psi)| \\
&\leq s(\|\phi + \psi\|^2 + \|\phi - \psi\|^2) \\
&= 2s(\|\phi\|^2 + \|\psi\|^2)
\end{aligned}$$

where $s = \sup_{\|\phi\|=1} |(T\phi, \phi)| = \sup_{\phi \neq 0} \frac{|(T\phi, \phi)|}{\|\phi\|^2}$. Let $\psi = tT\phi$, $t > 0$, so that the above becomes

$$\|T\phi\|^2 \leq \frac{s}{2}(\frac{1}{t}\|\phi\|^2 + t\|T\phi\|^2).$$

Take $t = \|\phi\|/\|T\phi\|$, so we find

$$\|T\phi\| \leq s\|\phi\|$$

as desired. □

Theorem 3.30. *(The spectral theorem.) Let H be a Hilbert space, let $T : H \to H$ a compact symmetric operator. Then:*

(1) Either $\|T\|$ or $-\|T\|$ is an eigenvalue for T.
(2) For each $r > 0$, the set $\{\lambda : \lambda$ is an eigenvalue for T and $|\lambda| > r\}$ is finite.
(3) For each eigenvalue $\lambda \neq 0$, $\ker(T - \lambda)$ is finite dimensional, i.e. each non-zero eigenvalue has finite geometric multiplicity.
(4) There exists an orthonormal basis $\{u_\alpha\}_{\alpha \in I}$ of H so that

$$Tu_\alpha = \lambda_\alpha u_\alpha.$$

All but countably many λ_α are zero.

Proof. (1) Let $\{\phi_n\}$ be a sequence of vectors so with $\|\phi_n\| = 1$ so that

$$\lim_{n\to\infty} |(T\phi_n, \phi_n)| = \|T\|;$$

by Lemma 3.29, such a sequence exists. As least one of $\|T\|$ or $-\|T\|$ must be an accumulation point of the sequence $(T\phi_n, \phi_n)$. We assume $\|T\|$ is - the other case proceeds nearly identically - and pass to a subsequence to assume that for the sequence ϕ_n, we have

$$\lim_{n\to\infty} (T\phi_n, \phi_n) = \|T\|.$$

By compactness of T, we can pass to a further subsequence to assume that $T\phi_n \to \psi$ for some $\psi \in H$. Now, consider

$$0 \leq (T\phi_n - \|T\|\phi_n, T\phi_n - \|T\|\phi_n) = \|T\phi_n\|^2 - 2\|T\|(\phi_n, T\phi_n) + \|T\|^2 \leq 2\|T\|^2 - 2\|T\|(\phi_n, T\phi_n)$$

and this right hand side goes to 0 as $n \to \infty$. It follows that $\|T\|\phi_n \to \psi$. If $\|T\| = 0$, then $T = 0$ and there is nothing to prove; if $\|T\| > 0$, then taking $\phi = \psi/\|T\|$, we find that $T\phi = \|T\|\phi$ and $\|\phi\| = 1$, as desired.

The rest of the claims in the spectral theorem now follow easily. To see (4), we use (1) to ensure the existence of an eigenvector u, then proceed by (transfinite) induction by replacing H with the subspace $H' = (\mathbb{R}u)^\perp$. We then note that $T(H') \subseteq H'$, and that $T|_{H'}$ is still symmetric and compact. To see (2) and (3), let $r > 0$ and let $H_r = \bigoplus_{|\lambda|>r} \ker(T - \lambda)$. Note that $T(H_r) \subseteq H_r$, and that the unit ball in this subspace $B_r = B \cap H_r$ has image $T(B_r)$ with compact closure. But by construction, $T(B_r) \supseteq rB_r$, so B_r is compact. This immediately implies H_r is finite dimensional. □

3.2.1. *Positive (semi-)definite operators.* Let $T : H \to H$ be a symmetric bounded operator.

Definition 3.31. If

$$(\phi, T\phi) \geq 0$$

for every $\phi \in H$, we say that T is *positive definite*.

Remark 3.32. It seems to be convention to call operators as above *positive definite*; it would be more accurate to call them *positive semi-definite*. We uneasily adopt the convention.

Remark 3.33. Observe that if T is positive definite and $u \in H$ is an eigenvector of T, i.e. if $Tu = \lambda u$, then $\lambda \geq 0$ as

$$0 \leq (u, Tu) = \lambda(u, u).$$

Lemma 3.34. *If T is a bounded operator, then T^*T is symmetric and positive definite.*

Proof. Note that

$$(\psi, T^*T\phi) = (T\psi, T\phi) = (T^*T\psi, \phi)$$

and that

$$(\phi, T^*T\phi) = (T\phi, T\phi) \geq 0.$$

□

Now let $T : H \to H$ be a compact operator, with adjoint T^*. Then T^*T is also compact, and by the above discussion, symmetric and positive definite. Similarly, TT^*T is also compact, symmetric, and positive definite.

Definition 3.35. For compact T, let $\{u_\alpha\}_{\alpha\in I}$ be an orthonormal eigenbasis of T^*T, with

$$T^*Tu_\alpha = \lambda_\alpha u_\alpha.$$

For each α let $\lambda_\alpha^{1/2}$ be the positive square root of λ_α, and let $|T| = (T^*T)^{1/2}$, i.e.

$$|T|u_\alpha := \lambda_\alpha^{1/2} u_\alpha.$$

$|T|$ is by construction a symmetric bounded positive definite operator.

Remark 3.36. We can remove the condition that T is compact at the cost of appealing to a more powerful version of the spectral theorem (for general bounded symmetric operators). This gives a definition of $|T|$ when T is a bounded operator. We will not need this construction in what follows, so we restrict our attention to the cheap definition above.

Remark 3.37. The observations above offer a version of the singular value decomposition for a compact operator T. As above, let $\{u_\alpha\}_{\alpha\in I}$ be an orthonormal eigenbasis for T^*T, with λ_α the corresponding eigenvalues. Observe that, for $\alpha, \beta \in I$,

$$(Tu_\alpha, Tu_\beta) = (T^*Tu_\alpha, u_\beta) = \lambda_\alpha(u_\alpha, u_\beta) = \begin{cases} \lambda_\alpha & \text{if } \alpha = \beta \\ 0 & \text{else} \end{cases}$$

so $Tu_\alpha = 0$ if and only if $\lambda_\alpha = 0$. Let $J = \{\alpha \in I : \lambda_\alpha > 0\} = \{\alpha \in I : Tu_\alpha \neq 0\}$, and let, for $\alpha \in J$, $v_\alpha := Tu_\alpha/\|Tu_\alpha\|$. Then $\{v_\alpha\}_{\alpha\in J}$ is a orthonormal set in H, and we may extend it to an orthonormal basis $\{v_\alpha\}_{\alpha\in I}$ of H with the same indexing set I.

Let $U : l^2(I) \to H$ be the isometry $(c_\alpha)_{\alpha\in I} \mapsto \sum_{\alpha\in I} c_\alpha u_\alpha$; similarly, let $V : l^2(I) \to H$ by given by $(c_\alpha)_{\alpha\in I} \mapsto \sum_{\alpha\in I} c_\alpha v_\alpha$. Let $\Sigma : l^2(I) \to l^2(I)$ be given by $(a_\alpha)_{\alpha\in I} \mapsto (\lambda_\alpha^{1/2} a_\alpha)_{\alpha\in I}$. This gives:

Corollary 3.38 (Singular value decomposition)**.** *Let $T : H \to H$ be compact. Let $U, V : l^2(I) \to H$, $\Sigma : l^2(I) \to l^2(I)$ be as above. Then*

$$T = V\Sigma U^*.$$

Remark 3.39. By truncating Σ, it is easy to use the above decomposition to show that every compact operator is the limit, in the operator norm, of finite rank operators. Thus $\mathrm{End}_{\mathrm{cm}}(H)$ is the closure of the space of finite rank operators in $\mathrm{End}(H)$.

If we note in addition that

$$|T| = U\Sigma U^*$$

then calling the unitary operator $W := VU^*$, we obtain

Corollary 3.40 (Polar decomposition)**.** *Let $T : H \to H$ be compact. Let $W : H \to H$, $\Sigma : l^2(I) \to l^2(I)$ be as above. Then*

$$T = W|T|$$

i.e. every compact operator can be written as a composition of a compact positive definite operator $|T|$ and a unitary operator W.

Corollary 3.41. *If T is compact, then T^* is as well.*

Proof. Write $T = W|T|$ with $|T|$ compact and positive definite. Then $T^* = |T|W^*$ is compact by Lemma 3.19. ◻

3.2.2. *Trace class operators.* We now use the spectral theorem and its corollaries to explain some basic properties of trace class operators on a Hilbert space.

Definition 3.42. Let H be a Hilbert space, $T : H \to H$ a compact positive definite operator. Let $\{e_\alpha\}_{\alpha\in I}$ be an orthonormal basis of H. We define

$$\operatorname{Tr}(T; \{e_\alpha\}_{\alpha\in I}) := \sum_{\alpha\in I}(e_\alpha, Te_\alpha).$$

Note that as $(e_\alpha, Te_\alpha) \geq 0$, this is a sum of non-negative real numbers hence the order of summation does not matter, but the sum may be $+\infty$.

Lemma 3.43. *Let $\{e_\alpha\}_{\alpha\in I}$ and $\{e'_\alpha\}_{\alpha\in I}$ be orthonormal bases of H.*

(i). If $T : H \to H$ is any bounded operator, then

$$\operatorname{Tr}(T^*T, \{e_\alpha\}_{\alpha\in I}) = \operatorname{Tr}(TT^*, \{e_\alpha\}_{\alpha\in I})$$

(ii). If $T : H \to H$ is compact, symmetric, and positive definite, and $U : H \to H$ is unitary, then

$$\operatorname{Tr}(UTU^*, \{e_\alpha\}_{\alpha\in I}) = \operatorname{Tr}(T, \{e_\alpha\}_{\alpha\in I})$$

(iii). If $T : H \to H$ is compact, symmetric, and positive definite, then

$$\operatorname{Tr}(T, \{e_\alpha\}_{\alpha\in I}) = \operatorname{Tr}(T, \{e'_\alpha\}_{\alpha\in I})$$

i.e. the trace is independent of the orthonormal basis used to compute it.

(iv). If T_1, T_2 are symmetric and positive definite bounded operators and $c \geq 0$, then

$$\operatorname{Tr}(T_1 + T_2, \{e_\alpha\}_{\alpha\in I}) = \operatorname{Tr}(T_1, \{e_\alpha\}_{\alpha\in I}) + \operatorname{Tr}(T_2, \{e_\alpha\}_{\alpha\in I})$$

and

$$\operatorname{Tr}(cT_1, \{e_\alpha\}_{\alpha\in I}) = c\operatorname{Tr}(T_1, \{e_\alpha\}_{\alpha\in I}).$$

Proof. (i). We have, appealing to the Pythagorean theorem and interchanging sums with non-negative entries

$$\begin{aligned}\operatorname{Tr}(T^*T, \{e_\alpha\}_{\alpha\in I}) &= \sum_{\alpha\in I}(e_\alpha, T^*Te_\alpha) = \sum_{\alpha\in I}(Te_\alpha, Te_\alpha) = \sum_{\alpha\in I}\|Te_\alpha\| \\ &= \sum_{\alpha\in I}\sum_{\beta\in I}(e_\beta, Te_\alpha)^2 = \sum_{\beta\in I}\sum_{\alpha\in I}(T * e_\beta, e_\alpha)^2 = \sum_{\beta\in I}\|T^*e\|^2 = \operatorname{Tr}(TT^*, \{e_\alpha\}_{\alpha\in I}).\end{aligned}$$

(ii). T is compact, symmetric, and positive definite, so let $\{u_\alpha\}_{\alpha\in I}$ be an orthonormal eigenbasis with eigenvalues $\lambda_\alpha \geq$. If we define an operator S by $Su_\alpha = \lambda_\alpha^{1/2}u_\alpha$, then S is bounded, symmetric, and $T = S^2 = S^*S$. We have $UTU^* = (US)(US)^*$, while $(US)^*(US) = S^2 = T$, so (ii). follows from (i). (iii). This is an immediate consequence of (ii) - simply let U^* be the unitary change-of-basis operator sending $U^*e_\alpha = e'_\alpha$ for all α. (iv). This is immediate from the definition, since all terms involved are non-negative, and so we can freely rearrange sums. □

Because of (iii). above, we simply write, for T compact, symmetric, and positive definite

$$\operatorname{Tr}(T) := \operatorname{Tr}(T, \{e_\alpha\}_{\alpha\in I})$$

for *any* choice of orthonormal basis.

Definition 3.44. Let T be a compact operator. We say T is *trace class* if $|T|$ satisfies

$$\operatorname{Tr}(|T|) < \infty.$$

We denote by $\operatorname{End}_1(H)$ the collection of trace class operators.

Lemma 3.45. *We have*

(i). Let T_1, T_2 be compact. Then

$$\operatorname{Tr}(|T_1 + T_2|) \leq \operatorname{Tr}(|T_1|) + \operatorname{Tr}(|T_2|).$$

(ii). Let T be compact, and S bounded. Then

$$\operatorname{Tr}(|TS|) \leq \operatorname{Tr}(|T|)\|S\|$$

and

$$\operatorname{Tr}(|ST|) \leq \|S\| \operatorname{Tr}(|T|)$$

Proof. (i). The polar decomposition (Corollary 3.40) allows us to write $T_1 = W_1|T_1|, T_2 = W_2|T_2|$, $T_1+T_2 = W|T_1+T_2|$. Then if $\{e_\alpha\}_{\alpha\in I}$ is an orthonormal basis, and $F \subseteq I$ is finite, we have

$$\sum_{\alpha\in F}(e_\alpha, |T_1+T_2|e_\alpha) = \sum_{\alpha\in F}(We_\alpha, (T_1+T_2)e_\alpha) = \left(\sum_{\alpha\in F}(W_1^*We_\alpha, |T_1|e_\alpha)\right) + \left(\sum_{\alpha\in F}(W_2^*We_\alpha, |T_2|e_\alpha)\right).$$

But, for $i = 1, 2$,

$$\begin{aligned}\sum_{\alpha\in F}|(W_i^*We_\alpha, |T_i|e_\alpha)| &= \sum_{\alpha\in F}|(|T_i|^{1/2}W_i^*We_\alpha, |T_i|^{1/2}e_\alpha)| \\ &\leq \sum_{\alpha\in F}\||T_i|^{1/2}W_i^*We_\alpha\|\||T_i|^{1/2}e_\alpha)\| \\ &\leq \left(\sum_{\alpha\in F}\||T_i|^{1/2}W_i^*We_\alpha\|^2\right)^{1/2}\left(\sum_{\alpha\in F}\||T_i|^{1/2}e_\alpha)\|^2\right)^{1/2}\end{aligned}$$

and since we have

$$\left(\sum_{\alpha\in F}\||T_i|^{1/2}e_\alpha)\|^2\right)^{1/2} = \left(\sum_{\alpha\in F}(e_\alpha, |T_i|e_\alpha)\right)^{1/2}$$

and, if $e'_\alpha = W_i^*We_\alpha$,

$$\left(\sum_{\alpha\in F}\||T_i|^{1/2}W_i^*We_\alpha\|^2\right)^{1/2} = \left(\sum_{\alpha\in F}(e'_\alpha, |T_i|e'_\alpha)\right)^{1/2}.$$

As F increases to I, these both tend to $\operatorname{Tr}(|T_i|)^{1/2}$, hence

$$\operatorname{Tr}(|T_1+T_2|) \leq \operatorname{Tr}(|T_1|) + \operatorname{Tr}(|T_2|).$$

(ii). By Lemma 3.19 ST and TS are both compact. We write the polar decompositions $T = W|T|$, $ST = W_1|ST|$, $TS = W_2|TS|$. Then

$$|ST| = (W_1^*SW)|T|$$

and since $\|W_1\| = \|W\| = 1$ we have $\|W_1^*SW\| \leq \|S\|$. Then

$$|ST|^2 = |T||(W_1^*SW)|^2|T|$$

so $\|W_1^*SW\||T|^2 - |ST|^2$ is positive definite, hence $\|S\|^2|T|^2 - |ST|^2$ is as well. We conclude that $\|S\||T| - |ST|$ is also positive definite, and so

$$\operatorname{Tr}(|ST|) \leq \|S\| \operatorname{Tr}(|T|).$$

For the other inequality, we simply note

$$\operatorname{Tr}(|TS|) = \operatorname{Tr}(|(TS)^*)| = \operatorname{Tr}(|S^*T^*|)$$

and that T^* is compact if T is by Corollary 3.41. □

Corollary 3.46. $\operatorname{End}_1(H)$ *is a two sided ideal of* $\operatorname{End}(H)$.

Remark 3.47. Indeed, $\operatorname{End}_1(H)$ is a $*$-ideal - it is stable under the taking the adjoint (obvious from definition). So we have nested inclusions of $*$-ideals:

$$\operatorname{End}_1(H) \subseteq \operatorname{End}_{\mathrm{cm}}(H) \subseteq \operatorname{End}(H)$$

inside the Banach algebra $\operatorname{End}(H)$. But be careful: $\operatorname{End}_1(H)$ is not closed in the operator norm on $\operatorname{End}(H)$, while $\operatorname{End}_{\mathrm{cm}}(H)$ is.

Proposition 3.48. *Let H be a Hilbert space. The following are equivalent:*

(i). T is trace class, i.e. $\operatorname{Tr}(|T|) < \infty$.

(ii). T is a finite linear combination of compact, symmetric, positive definite operators with finite trace.

(iii). $\sum_{\alpha\in I} |(e'_\alpha, Te_\alpha)| < \infty$ *for any choice of two orthonormal bases* $\{e_\alpha\}_{\alpha\in I}, \{e'_\alpha\}_{\alpha\in I}$.
(iv). $\sum_{\alpha\in I} |(e'_\alpha, Te_\alpha)| < \infty$ *for any one fixed orthonormal basis* $\{e_\alpha\}_{\alpha\in I}$ *and all other orthonormal bases* $\{e'_\alpha\}_{\alpha\in I}$.
(v). $\sum_{\alpha\in I} |(e_\alpha, Te_\alpha)| < \infty$ *for every orthonormal basis* $\{e_\alpha\}_{\alpha\in I}$.

Proof. [(i) $\iff$ (ii).] (ii). $\implies$ (i). is apparent from Corollary 3.46. (i). $\implies$ (ii). since, given T of trace class, we can let

$$T_{\text{sym}} := \frac{1}{2}(T+T^*), T_{\text{sk}-\text{sym}} := \frac{1}{2}(T-T^*)$$

and note that T_{sym} is compact and symmetric, while $T_{\text{sk}-\text{sym}}$ is compact and skew symmetric, i.e. $T^*_{\text{sk}-\text{sym}} = -T_{\text{sk}-\text{sym}}$. Since $\frac{1}{i}T_{\text{sk}-\text{skew}}$ is a Hermitian operator, we may therefore reduce to the case that T is itself compact and symmetric.[2] Now, if T is compact symmetric, let $\{u_\alpha\}_{\alpha\in I}$ be an eigenbasis, and write $T = U\Sigma U^*$ where $U : l^2(I) \to H$ is the unitary operator $U((c_\alpha)_{\alpha\in I}) = \sum_{\alpha\in I} c_\alpha u_\alpha$, and where $\Sigma : l^2(I) \to l^2(I)$ satisfies $\Sigma((u_\alpha)_{\alpha\in I}) = (\lambda_\alpha u_\alpha)_{\alpha\in I}$. Then let $T_+ = U\Sigma_+ U^*, T_+ = U\Sigma_- U^*$ where Σ_+ consists of just the positive eigenvalues and Σ_- consists of the negatives of the negative eigenvalues. Thus

$$T = T_+ - T_-, |T| = T_+ + T_-.$$

and T_+, T_- are both compact and positive definite. It follows that every trace class operator can be written as a linear combination of 4 compact positive definite operators with finite trace.

[(ii). $\implies$ (iii).] Let $\{e_\alpha\}_\alpha, \{e'_\alpha\}_\alpha$ be orthonormal bases, and let U be the unitary change of basis $Ue'_\alpha = e_\alpha$. Then UT is also trace class by Lemma 3.46, and by (ii). we write $UT = \sum_{k=1}^K \rho_k P_k$, where each P_k is compact and positive definite with finite trace. Then

$$|(e'_\alpha, Te_\alpha)| = |(e_\alpha, UTe_\alpha)| \leq \sum_{k=1}^K |\rho_k|(e_\alpha, P_k e_k) < \infty$$

[(iii). $\implies$ (iv).] Apparent.

[(iii). $\implies$ (v).] Apparent.

[(iv). $\implies$ (i).] Let $\{e_\alpha\}_{\alpha\in I}$ be any one fixed orthonormal basis, and suppose that for all orthonormal bases $\{e'_\alpha\}_{\alpha\in I}$,

$$\sum_{\alpha\in I} |(e'_\alpha, Te_\alpha)| < \infty.$$

We use Corollary 3.40 to write $T = W|T|$. Then if we choose $e'_\alpha = We_\alpha$ for all α

$$\text{Tr}(|T|) = \sum_{\alpha\in I} |(W^* e'_\alpha, |T|e_\alpha)| = \sum_{\alpha\in I} |(e'_\alpha, Te_\alpha)| < \infty.$$

[(v). $\implies$ (i).] If T satisfies (v). then so does T^*, hence so do $T_{\text{sym}} := \frac{1}{2}(T+T^*)$ and $T_{\text{sk}-\text{sym}} := \frac{1}{2}(T-T^*)$. So let us assume that T is not just bounded, but symmetric. Now, note that for any infinite sequence of distinct elements $\{\alpha_k\}_{k=1}^\infty \subseteq I$, we have

$$\lim_{k\to\infty} |(e_{\alpha_k}, Te_{\alpha_k})| \to 0.$$

We claim, without proof, that this together with the fact that T is symmetric, is enough to deduce that T is compact.[3] Given this black box, we are done: the symmetric compact T is diagonalizable, and applying (v) to its eigenbasis allows us to conclude. □

Definition 3.49. Let T be trace class. Then

$$\text{Tr}(T) := \sum_{\alpha\in I} (e_\alpha, Te_\alpha)$$

and this sum is absolutely convergent and independent of choice of orthonormal basis $\{e_\alpha\}_{\alpha\in I}$ used to compute it.

Lemma 3.50. *Let H be a Hilbert space, $T \in \text{End}(H)$. If T can be written as the composition of two Hilbert-Schmidt operators, then T is trace-class.*

[2]OK, strictly speaking we have to work with Hermitian operators, i.e. complex values Hilbert spaces - but we will ignore this easily fixed inaccuracy.

[3]For a reference, see https://math.stackexchange.com/questions/66329/a-question-on-compact-operators.

Proof. Let $T = B \circ A$ for some $A, B \in \mathrm{HS}(H)$. Let $\{e_\alpha\}_{\alpha \in I}$ be an orthonormal basis of H, and let $a_{\alpha,\beta} \in \mathbb{R}$ be the "matrix coefficients" of A, i.e.

$$Ae_\alpha = \sum_{\beta \in I} a_{\alpha,\beta} e_\beta.$$

Similarly, let $b_{\alpha,\beta}$ be the matrix coefficients of B, and $c_{\alpha,\beta}$ those of T. Clearly

$$c_{\alpha,\beta} = \sum_{\gamma \in I} b_{\alpha,\gamma} a_{\gamma,\beta}$$

and we find that by Cauchy-Schwartz,

$$\sum_\alpha |c_{\alpha,\alpha}| = \sum_\alpha \sum_\gamma b_{\alpha,\gamma} a_{\gamma,\alpha} \leq \left(\sum_{\alpha,\gamma} b_{\alpha,\gamma}^2 \right)^{1/2} \left(\sum_{\alpha,\gamma} a_{\gamma,\alpha}^2 \right)^{1/2} = \|B\|_{\mathrm{HS}(H)} \|A\|_{\mathrm{HS}(H)} < \infty.$$

□

Remark 3.51. The converse (every trace class operator is the composition of tw Hilbert-Schmidt operators) holds as well, but we omit its proof as we will not need it.

Remark 3.52. Trace class operators play a special role in the definition of the Hilbert-Schmidt norm. Namely, we have mentioned that $\mathrm{HS}(H) = \widehat{a}(H^\vee \widehat{\otimes} H)$ inherits an inner product. What is this on the level of endomorphisms? Well, simply tracing through the definitions, we can find that the $H^\vee \widehat{\otimes} H$ inner product becomes

$$(T, S)_{\mathrm{HS}(H)} := \mathrm{Tr}(T^* S).$$

3.3. **Mercer's theorem, measurable version.** Let $K \in L^2_{\mathrm{HS}}(\mathcal{X} \times \mathcal{X})$ be a symmetric Hilbert-Schmidt kernel on $\mathcal{X}$.

Definition 3.53. We say that K satisfies *Mercer's condition* if T_K is positive definite, i.e.

$$(\phi, T_K \phi) = \int_{\mathcal{X}} \int_{\mathcal{X}} \phi(x) K(x,y) \phi(y) dx dy \geq 0$$

for every square integrable measurable ϕ.

Definition 3.54. We call a symmetric positive definite Hilbert-Schmidt kernel on $\mathcal{X}$ a *Mercer kernel.*[4]

Lemma 3.55. *Let K be a Mercer kernel on $\mathcal{X}$. If λ is an eigenvalue of T_K, then λ is non-negative.*

Proof. Let u be an eigenvector with eigenvalue λ. By Lemma 3.28, λ must be real. Since $0 \leq (u, T_K u) = \lambda(u,u)$, λ is non-negative. □

Let K be a Mercer kernel on $\mathcal{X}$, and let $\lambda_1, \lambda_2, \ldots$ be the non-zero eigenvalues of T_K (there are only at most countably many such), counted with multiplicity and ordered in descending order, with corresponding orthonormal set of eigenvectors $u_1, u_2, \ldots$.

Theorem 3.56 (Mercer's theorem, measurable version)**.** *K can be expanded as*

$$K(x,y) = \sum_{n=1}^\infty \lambda_n u_n(x) u_n(y)$$

with convergence in the L^2 norm.

Proof. There is nothing left to prove! Simply complete the orthonormal collection $\{u_n\}_{n=1}^\infty$ to an orthonormal basis of $L^2(\mathcal{X})$. Then the theorem follows from Proposition 3.14. □

[4]This may not be standard terminology, but I got tired of repeating the same adjectives.

4. Mercer's theorem: continuous form

The contents of this section are heavily indebted to [MHC12]. [5]

We move now to the continuous version of Mercer's theorem. This result builds on the measurable version of the previous section by imposing relatively mild continuity conditions on a kernel K, and obtaining as a result a pointwise version of Mercer's decomposition $K(x,y) = \sum \lambda_n e_n(x) e_n(y)$. It is also this version that admits a satisfying interpretation in terms of existence of a feature map.

We start with some generalities. Let X be a topological space, let $\mathfrak{B}$ be the Borel σ-algebra, and let μ be a measure on $(X, \mathfrak{B})$.

Definition 4.1. μ is *strictly positive* if $\mu(U) > 0$ whenever U is open and non-empty.

Remark 4.2. Recall that the support of a Borel measure μ is defined by

$$\operatorname{supp}(\mu) := \{x \in X : \text{for all open } U \ni x, \mu(U) > 0\}.$$

It is easy to see that $\operatorname{supp}(\mu) = X$ if and only if μ is strictly positive.

Definition 4.3. μ is *locally finite* if, for every $x \in X$, there is an open U containing x with $\mu(U) < \infty$.

For the remainder of this section, we will assume that μ is strictly positive, locally finite, and σ-finite.

Remark 4.4. Note that:

(1) X need not be compact, or even locally compact;
(2) X does not need to satisfy any separability or countability axiom;
(3) μ need not be regular;
(4) μ need not be finite.

4.1. Smooth kernels.

Let K be a Mercer kernel on X.

Definition 4.5. We say that a Mercer kernel K is *a smooth kernel* if

(i). K is continuous (as a function on $X \times X$) at every point of ΔX;
(ii). For every $x \in X$, $K(x, \cdot)$ is square integrable on X;
(iii). The map

$$\begin{aligned} X &\to L^2(X) \\ x &\mapsto K(x, \cdot) \end{aligned}$$

is continuous.

4.1.1. *First observations.* The conditions (i).,(ii).,(iii)., imposed above can all be easily explained. We ask for (i)., continuity along the diagonal, because of the following simple observation.

Lemma 4.6. *Let K be a Mercer kernel which is continuous at every point of the diagonal $\Delta X \subseteq X \times X$. Then $K(x,x) \geq 0$ for all $x \in X$.*

Proof. Let $x_0 \in X$ be such that $K(x_0, x_0) < 0$. Let $\varepsilon > 0$ be such that $K(x_0, x_0) < -\varepsilon < 0$. By continuity, there exists non-empty open $U \ni x_0$ so that for all $x, y \in U$,

$$|K(x,y) - K(x_0, x_0)| < \frac{\varepsilon}{2}.$$

By strict positivity and local finiteness of μ, we may also assume $0 < \mu(U) < \infty$. Now, $K(x,y) < -\varepsilon/2$ on $U \times U$. If we call $\phi = \mathbb{1}_U$, we find that

$$\int_X \int_X \phi(x) K(x,y) \phi(y) dx dy < -\frac{\varepsilon \mu(U)^2}{2} < 0$$

which contradicts positive definiteness. □

(ii)., the square integrability of the slices $K(x, \cdot)$, allows us to make sense of the requirement (iii).; (iii). is useful because of the following.

[5] Unfortunately, we believe that [MHC12] has a few faulty arguments - see their Theorem 2.7 in particular. For that reason, we have rewritten much of their exposition, and, in a few places, made stronger assumptions to ensure correctness.

Lemma 4.7. *Let K on $X \times X$ be a smooth kernel.*

1. *Let ϕ be a square integrable function on X. Then $T_K\phi$ is continuous on X.*
2. *If ϕ_1, ϕ_2 are square integrable and $\phi_1(y) = \phi_2(y)$ for a.e. $y \in X$, then $T_K\phi_1(x) = T_K\phi_2(x)$ for all $x \in X$.*

We thus say that, if $\phi \in L^2(X)$, then $T_K\phi$ is continuous.

Proof. (1) By definition

$$T_K\phi(x) = (K(x, \cdot), \phi).$$

Taking the inner product with ϕ is a continuous linear functional, and the map $x \mapsto K(x, \cdot)$ is continuous in x by condition (iii) in the definition of a smooth kernel. The composition is therefore continuous. For (2), note that the classes of ϕ_1 and ϕ_2 in $L^2(X)$ are the same. Thus $T_K\phi_1$ and $T_K\phi_2$ represent the same class in $L^2(X)$, i.e. are equal almost everywhere. By (1), they are both continuous - appealing to the following (easy) observation concludes. $\square$

Lemma 4.8. *Let ψ_1, ψ_2 be continuous functions on X with $\psi_1(x) = \psi_2(x)$ for a.e. x. Then $\psi_1 = \psi_2$ identically.*

Proof. We proceed by contradiction. Let x_0 be such that $\psi_1(x_0) \neq \psi_2(x_0)$, and let ε satisfy $0 < \varepsilon < |\psi_1(x_0) - \psi_2(x_0)|/2$. Then there exists open neighborhoods U_1 and U_2 of x_0, so that $|\psi_i(x) - \psi_i(x_0)| < \varepsilon$ on U_i. Hence on $U = U_1 \cap U_2$, $\psi_1(x) \neq \psi_2(x)$. Since μ is strictly positive and $U \neq \emptyset$, $\mu(U) > 0$. $\square$

Corollary 4.9. *If $u \in L^2(X)$ satisfies $T_K u = \lambda u$ with $\lambda > 0$, the u has a representative which is continuous on X.*

4.1.2. *Mercer's theorem, smooth kernels version.* By Theorem 3.56 (the measurable form of Mercer's theorem), we see that for a smooth kernel, we can write, in L^2

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y) \tag{4.1}$$

where the (countably many) positive eigenvalues $\lambda_n > 0$ are arranged in decreasing order, and u_n are orthonormal and *continuous*.

Our goal in this section is to upgrade this identity in two ways. We want:

1. equality in (4.1) to hold pointwise everywhere, and
2. absolute and uniform convergence of the series $\sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y)$ on compact subsets of $X \times X$.

Note that if we can ensure (1) and (2), and if we impose in addition that X be locally compact, then $K(x, y)$ is automatically continuous on $X \times X$. This also makes it clear that (1) and (2) cannot follow from K being smooth alone. For instance, if K is a smooth kernel, then changing the value of K on an appropriate null set in $X \times X - \Delta X$ will keep K smooth - this is incompatible with (1) and (2), which would imply that K is continuous.

Still, let's see how far we can get with just the requirement that K is a smooth kernel.

The next result explains how, after restricting to the diagonal, the two sides of (4.1) can be seen to relate to one another.

Proposition 4.10. *Let K be a smooth kernel. Then*

$$\sum_{n=1}^{\infty} \lambda_n u_n(x)^2 \leq K(x, x)$$

Proof. First, note that for any $\phi \in L^2(X)$,

$$0 \leq \sum_{n=1}^{\infty} \lambda_n (\phi, u_n)^2 < \infty$$

since this expression is simply $(\phi, T_K\phi)$.

Let N be a fixed positive integer. Consider the "tail end of the series"

$$K_N(x, y) := K(x, y) - \sum_{n=1}^{N} \lambda_n u_n(x) u_n(y).$$

As the difference of Hilbert-Schmidt symmetric kernels, this is visibly Hilbert-Schmidt and symmetric. By Corollary 4.9 the u_n are all continuous, so K_N is also continuous at every point of the diagonal.

We claim that K_N is positive definite (hence a Mercer kernel). To see this, write

$$\int_X \int_X \phi(x) K_N(x,y)\phi(y)dxdy = \int_X \phi(x)\left(T_K\phi(x) - \sum_{n=1}^{N} \lambda_n (u_n,\phi)u_n(x)\right)dx = \sum_{n=N+1}^{\infty} \lambda_n (u_n,\phi)^2 \geq 0.$$

It follows that K_N satisfies the conditions of Lemma 4.6, and so $K_N(x,x) \geq 0$ for any $x \in X$. But then for each N,

$$\sum_{n=1}^{N} \lambda_n u_n(x)^2 \leq K(x,x).$$

Since N was arbitrary, we can conclude. □

We can use this to immediately ensure a weak form of (2), the desired pointwise convergence of the series $\sum_{n=1}^{\infty} \lambda_n u_n(x)u_n(y)$.

Proposition 4.11. *Let K be a smooth kernel. Then the series*

$$\sum_{n=1}^{\infty} \lambda_n u_n(x)u_n(y)$$

is absolutely and uniformly convergent on compact subsets of X with respect to each variable, when the other is fixed.

Proof. Fix $y \in X$ and let $\Xi \subseteq X$ be compact. For $x \in \Xi$, and $M \geq N \geq 1$, we have by Cauchy-Schwartz and Proposition 4.10

$$\begin{aligned}\left(\sum_{n=N}^{M} |\lambda_n u_n(x)u_n(y)|\right)^2 &= \left(\sum_{n=N}^{M} (\lambda_n^{1/2}|u_n(x)|)(\lambda_n^{1/2}|u_n(y)|)\right)^2) \\ &\leq \left(\sum_{n=1}^{\infty} \lambda_n u_n(x)^2\right)\left(\sum_{n=N}^{M} \lambda_n u_n(y)^2\right) \\ &\leq \sup_{\xi\in\Xi} K(\xi,\xi) \sum_{n=N}^{M} \lambda_n u_n(y)^2.\end{aligned}$$

By the continuity of $K(x,y)$ at the diagonal, we have $\sup_{\xi\in\Xi} K(\xi,\xi) < \infty$. Since $\sum_{n=1}^{\infty} \lambda_n u_n(y)^2 \leq K(y,y) < \infty$, it follows immediately that for fixed y, the series $\sum_{n=1}^{\infty} \lambda_n u_n(x)u_n(y)$ converges absolutely and uniformly on $x \in \Xi$. □

Proposition 4.12. *Suppose X is locally compact. Let K be a smooth kernel, and suppose that for every x,*

$$K(x,\cdot) : y \mapsto K(x,y)$$

is continuous on X. Then the partial sums of

$$\sum_{n=1}^{\infty} \lambda_n u_n(x)^2$$

are all continuous, non-negative, and monotonically converge to $K(x,x)$ for every x.

Proof. Continuity, non-negativity, and monotonicity are apparent - what we must prove is that, for every x, $K(x,x) = \sum_{n=1}^{\infty} \lambda_n u_n(x)^2$.

Let $x \in X$. We have, since $\lambda_n \leq \|T\|$ for all n,

$$\sum_{n=1}^{\infty}(\lambda_n u_n(x))^2 \leq \|T\| \sum_{n=1}^{\infty} \lambda_n u_n(x)^2 \leq \|T\| K(x,x)$$

by Proposition 4.10. If follows that for each $x \in X$, the series

$$K_x := \sum_{n=1}^{\infty} \lambda_n u_n(x)u_n$$

converges in $L^2(X)$. Note that

$$\int_X K_x(y)\phi(y)dy = T_K\phi(x)$$

for all $x \in X$ and $\phi \in L^2(X)$, so we have that for every x, $K_x(y) = K(x,y)$ for almost every y. Given $y_0 \in X$, by local compactness of X there exists a compact set Ξ containing an open neighborhood of y_0. By Proposition 4.11, for each fixed x, $\sum_{n=1}^\infty \lambda_n u_n(x)u_n(y)$ converges uniformly to $K_x(y)$ on $y \in \Xi$ - thus, $K_x(\cdot)$ is continuous at y_0; since y_0 was arbitrary, $K_x(\cdot)$ is continuous everywhere. Since $K(x,\cdot)$ is also continuous and we have $K_x(y) = K(x,y)$ for almost every y, Lemma 4.8 forces equality for all y. It follows that

$$K(x,x) = K_x(x) = \sum_{n=1}^\infty \lambda_n u_n(x)^2$$

for every $x \in X$. □

Corollary 4.13. *Under the same assumptions as above, the series $\sum_{n=1}^\infty \lambda_n u_n(x)^2$ converges uniformly on compact sets to $K(x,x)$.*

Proof. The monotonicity of the partial sums, together with the fact that their pointwise limit $K(x,x)$ is continuous, allows us to simply appeal to Dini's theorem to conclude. □

Proposition 4.14. *Suppose X is locally compact. Let K be a smooth kernel, and suppose that for each x, $K(x,\cdot)$ is continuous. Then the series*

$$\sum_{n=1}^\infty \lambda_n u_n(x)u_n(y)$$

converges absolutely and uniformly on compact subsets of $X \times X$.

Proof. Let $C \subseteq X \times X$ be a compact set. For each $(x,y) \in C$, there exists, by local compactness of X, open neighborhoods U_x, U_y of x and y, respectively, with compact closures. $\{U_x \times U_y\}_{(x,y)\in C}$ is an open cover of the compact C, hence there is a finite subcover $\{U_{x_i} \times U_{y_i}\}_{i=1}^N$.

Thus, it suffices to show that for $U, V \subseteq X$ with compact closure, that $\sum_{n=1}^\infty \lambda_n u_n(x)u_n(y)$ is absolutely and uniformly convergent on $\overline{U} \times \overline{V}$.

Let $1 \le N \le M$; again we have by Cauchy-Schwartz

$$\left(\sum_{n=N}^M \lambda_n |u_n(x)u_n(y)|\right)^2 \le \left(\sum_{n=N}^M \lambda_n u_n(x)^2\right)\left(\sum_{n=N}^M \lambda_n u_n(y)^2\right)$$

Since, by Corollary 4.13, $\sum_{n=1}^\infty \lambda_n u_n(x)^2$ converges uniformly to $K(x,x)$ on $\overline{U}$ and $\sum_{n=1}^\infty \lambda_n u_n(y)^2$ converges uniformly to $K(y,y)$ on $\overline{V}$, the sums $\sum_{n=N}^M \lambda_n u_n(x)^2$ and $\sum_{n=N}^M \lambda_n u_n(y)^2$ can be made arbitrarily small uniformly for $(x,y) \in \overline{U} \times \overline{V}$. □

Theorem 4.15 (Mercer's theorem, smooth kernel version)**.** *Suppose X is locally compact. Let K be a smooth kernel, and suppose that for every $x \in X$, $K(x,\cdot)$ is continuous. Then for every $(x,y) \in X \times X$,*

$$K(x,y) = \sum_{n=1}^\infty \lambda_n u_n(x)u_n(y)$$

and this series converges absolutely and uniformly on compact subsets of $X \times X$.

Proof. There is almost nothing left to prove - we have already shown that the series converges absolutely and uniformly on compacts. If we set $\widetilde{K}(x,y) = \sum_{n=1}^\infty \lambda_n u_n(x)u_n(y)$, then by uniform convergence, $\widetilde{K}$ is continuous on $X \times X$. By construction,

$$T_K = T_{\widetilde{K}}$$

so $K = \widetilde{K}$ a.e. on $X \times X$. But by Lemma 4.8 applied to $X \times X$, $K(x,y) = \widetilde{K}(x,y)$ for all $x, y \in X$. □

4.2. **Smooth kernels occur in the wild.** Thus far, our discussion of the continuous form of Mercer's theorem has relied on the artificial notion of a smooth kernel. In this subsection, we will show that, in most situations, the conditions (i),(ii), and (iii) in the definition of a smooth kernel are automatic.

We start with the "classical" setting for Mercer theory.

Lemma 4.16. *Suppose X is a compact metric space, μ a probability Borel measure on X whose support is X. Then if K is continuous on $X \times X$, symmetric and positive definite, then K is a smooth kernel.*

Proof. Asking that $\mathrm{supp}(\mu) = X$ is just another way of saying that μ is strictly positive. μ is also locally finite, as it is a probability Borel measure. Thus our running assumptions on μ are satisfied.

(Just for psychological ease, one can note further that μ, as a probability Borel measure on a compact metric space, is automatically regular - we will not need this, but it is reassuring.)

Let us see that K is a smooth kernel. Condition (i). is apparent. Every continuous function on X is square-integrable as $\mu(X) = 1$, so (ii). holds as well. It remains to see (iii). But K is continuous on $X \times X$, hence uniformly continuous. Thus, given any $\varepsilon > 0$, there exists $\delta > 0$ so that if $d_{X\times X}((x, y), (x', y')) < \delta$,

$$|K(x, y) - K(x', y')| < \varepsilon.$$

But then, if $d(x, y) < \delta$,

$$\|K(x, \cdot) - K(y, \cdot)\|^2 = \int_X |K(x, z) - K(y, z)|^2 d\mu(z) \leq \varepsilon^2 \int_X d\mu(z) = \varepsilon^2.$$

□

This can be easily generalized. In what follows, X is an arbitrary topological space, and μ a Borel measure which is strictly positive, locally finite, and σ-finite.

Proposition 4.17. *Let $K(x, y)$ be a continuous Mercer kernel on $X \times X$. Then*

$$K(x, y)^2 \leq K(x, x)K(y, y)$$

for every $x, y \in X$.

Remark 4.18. Compare this result to the Cauchy-Schwartz identity

$$\left(\sum_{n=1}^{\infty} \lambda_n u_n(x)u_n(y)\right)^2 \leq \left(\sum_{n=1}^{\infty} \lambda_n u_n(x)^2\right)\left(\sum_{n=1}^{\infty} \lambda_n u_n(y)^2\right)$$

underlying the proof of Proposition 4.11.

Proof. Let $(x_0, y_0) \in X \times X$. If $K(x_0, y_0) = 0$, then there is nothing to prove. Else, fix ε with $0 < \varepsilon < |K(x_0, y_0)|$. By continuity of K, there exists opens $U, V \subseteq X$ with $x_0 \in U, y_0 \in V$, so that

$$|K(x, x') - K(x_0, x_0)| < \varepsilon \text{ if } x, x' \in U,$$

$$|K(y, y') - K(y_0, y_0)| < \varepsilon \text{ if } y, y' \in V,$$

$$|K(x, y) - K(x_0, y_0)| < \varepsilon \text{ if } x \in U, y \in V.$$

Since X is locally finite, we may assume that $\mu(U), \mu(V) < \infty$. Let $\delta_U = \frac{1}{\mu(U)}\mathbb{1}_U$, $\delta_V = \frac{1}{\mu(V)}\mathbb{1}_V$, and consider, for each $t \in [0, 1]$,

$$\phi_t = t\delta_U + \delta_V.$$

Consider the quadratic function of t

$$q(t) := (\phi_t, T_K\phi_t) = (\delta_U, T_K\delta_U)t^2 + 2(\delta_U, T_K\delta_V)t + (\delta_V, T_K\delta_V)$$

By positive definiteness of the kernel K,

$$q(t) \geq 0,$$

hence the discriminant Δ of $q(t)$ must be less than or equal to 0. In other words,

$$\Delta = 4(\delta_U, T_K\delta_V)^2 - 4(\delta_U, T_K\delta_U)(\delta_V, T_K\delta_V) \leq 0$$

i.e.,

$$(\delta_U, T_K\delta_V)^2 \leq (\delta_U, T_K\delta_U)(\delta_V, T_K\delta_V).$$

By our choices of U and V, we have

$$|(\delta_U, T_K\delta_U) - K(x_0, x_0)| < \varepsilon$$

$$|(\delta_V, T_K\delta_V) - K(y_0, y_0)| < \varepsilon$$

$$|(\delta_U, T_K\delta_V) - K(x_0, y_0)| < \varepsilon.$$

Taking a limit as $\varepsilon \to 0$ gives

$$K(x_0, y_0)^2 \leq K(x_0, x_0)K(y_0, y_0).$$

□

Recall the following notion from point set topology.

Definition 4.19. A topological space X is called *first countable* if every point of X has a countable neighborhood base.

This very weak countability axiom is useful because of the following fact.

Lemma 4.20. *Let X be a topological space. Suppose X is first countable, then a function $f : X \to \mathbb{R}$ is continuous if and only if, for every sequence $x_n \to x$, $f(x_n) \to f(x)$.*

Proof. Left as an easy exercise. □

Proposition 4.21. *Let X be a first countable topological space, equipped with μ a non-degenerate, locally finite, and σ-finite Borel measure. Let K be a continuous Mercer kernel, and suppose too that*

$$\int_X K(x, x)dx < \infty.$$

Then K is a smooth kernel.

Proof. By assumption, (i) in the definition of a smooth kernel is automatic.

By Proposition 4.2, we have, for every $x \in X$.

$$\int_X K(x, y)^2 dy \leq K(x, x) \int_X K(y, y)dy < \infty$$

so (ii) is also immediate.

To see (iii), the continuity of $x \mapsto K(x, \cdot)$, let $x_0 \in X$. Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in X tending towards x_0. We have

$$\begin{aligned}(K(x_n, y) - K(x_0, y))^2 &= 2K(x_n, y)^2 + 2K(x_0, y)^2 - (K(x_n, y) + K(x_0, y))^2 \\ &\leq 2K(x_n, y)^2 + 2K(x_0, y)^2 \\ &\leq 2K(y, y)(K(x_n, x_n) + K(x_0, x_0))\end{aligned}$$

by Proposition 4.2. By continuity of K, since $x_n \to x_0$, we have that $M = \sup_n K(x_n, x_n) < \infty$. Thus, we can bound $y \mapsto (K(x_n, y), K(x_0, y))^2$ by the integrable function $2(M + K(x_0, x_0))K(y, y)$. Since $(K(x_n, y) - K(x_0, y))^2$ as $n \to \infty$ for each y, we can apply the dominated convergence theorem[6] to conclude that

$$\|K(x_n, \cdot) - K(x_0, \cdot)\|_{L^2(X)}^2 = \int_X (K(x_n, y) - K(x_0, y))^2 dy \to 0.$$

Since X is first countable, Lemma 4.20 implies that $x \mapsto K(x, \cdot)$ is continuous. □

[6]There is no version (and there cannot be a version) of the DCT for nets, so we must pass to a countable sequence $\{x_n\}_n$ to run this argument.

4.2.1. *Mercer's theorem, continuous form.* We assemble all of our discussion thus far in the following.

Theorem 4.22 (Mercer's theorem, continuous form)**.** *Let X be locally compact and first countable. Suppose that K is a continuous Mercer kernel satisfying*

$$\int_X K(x,x)dx < \infty.$$

Then for every $(x,y) \in X \times X$,

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y)$$

and this series converges absolutely and uniformly on compact subsets of $X \times X$.

Proof. Simply combine Proposition 4.21 with Theorem 4.15. □

Proposition 4.23. *Let X, μ, K be as in the continuous version Mercer's theorem above. Then T_K is trace class, and*

$$\mathrm{Tr}(T_K) = \sum_{n=1}^{\infty} \lambda_n = \int_X K(x,x)dx.$$

Proof. Since T_K is a positive compact operator, the sum

$$\mathrm{Tr}(T_K; \{e_\alpha\}_{\alpha \in I} := \sum_{\alpha \in I} (e_\alpha, T_K e_\alpha)$$

consists entirely of positive terms and, as an element of $[0,\infty]$, is independent of choice of orthonormal basis $\{e_\alpha\}_\alpha$.

So let $\{u_1, u_2, \dots\}$ be the orthonormal set of eigenfunctions with positive eigenvalues, and extends this set to an orthonormal basis $\{u_\alpha\}_{\alpha \in I}$ of $L^2(X)$. Then

$$\mathrm{Tr}(T_K; \{u_\alpha\}_{\alpha \in I} = \sum_{n=1}^{\infty} \lambda_n.$$

On the other hand, the sum

$$\sum_{i=1}^{\infty} \lambda_n u_n(x)^2$$

converges monotonically to $K(x,x)$ by Proposition 4.12, so by the monotone convergence theorem,

$$\sum_{n=1}^{\infty} \lambda_n = \int_X K(x,x)dx < \infty.$$

Thus,

$$\mathrm{Tr}(T_K; \{e_\alpha\}_{\alpha \in I}$$

is finite for every orthonormal basis, and is independent of orthonormal basis, i.e. T_K is trace class. □

4.3. **The feature map interpretation.** For this subsection, we assume the hypotheses of Theorem 4.22, i.e. that X is locally compact and first countable; μ is strictly positive, locally finite, and σ-finite; and K is a continuous Mercer kernel with

$$\int_X K(x,x)dx < \infty.$$

Recall the following basic facts, established in the previous subsections.

(1) The eigenvectors of T_K in $L^2(X)$ all have continuous representatives u_n, so there is no ambiguity in choosing representatives of vectors in L^2 – there is a best choice u_n.
(2) The series $\sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y)$ converges absolutely and uniformly to $K(x,y)$ everywhere. In particular, we do not have to limit ourselves to a.e. pointwise convergence, and we do not need to pass to a subsequence of partial sums to get a pointwise convergence.

With these remarks in mind, note the following.

Lemma 4.24. *For each $x \in X$, we have*

$$(\lambda_n^{1/2} u_n(x))_{n=1}^{\infty} \in l^2(\mathbb{N}).$$

Thus, for every x,

$$\sum_{n=1}^{\infty} \lambda_n^{1/2} u_n(x) u_n \in L^2(X).$$

Proof. We have, by Proposition 4.12,

$$\sum_{n=1}^{\infty} (\lambda_n^{1/2} u_n(x))^2 = \sum_{n=1}^{\infty} \lambda_n u_n(x)^2 = K(x,x) < \infty$$

for every $x \in X$. □

Definition 4.25. The *feature map* associated with K

$$\Phi_K : X \to L^2(X)$$

$$x \mapsto \Phi_K(x) := \sum_{n=1}^{\infty} \lambda_n^{1/2} u_n(x) u_n.$$

Proposition 4.26. *The feature map $\Phi_K : X \to L^2(X)$ is continuous.*

Proof. We can simply compute that, if $x_0 \in X$,

$$\|\Phi_K(x) - \Phi_K(x_0)\|_{L^2(X)} = \sum_{n=1}^{\infty} \lambda_n (u_n(x) - u_n(x_0))^2 = \left(\sum_{n=1}^{\infty} \lambda_n u_n(x)^2 \right) + \left(\sum_{n=1}^{\infty} \lambda_n u_n(x_0)^2 \right) - 2 \left(\sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(x_0) \right).$$

Hence, by Mercer's theorem,

$$\|\Phi_K(x) - \Phi_K(x_0)\|_{L^2(X)} = K(x,x) + K(x_0,x_0) - 2K(x,x_0).$$

We conclude after appealing to the continuity of K. □

The feature map Φ_K is designed to satisfy the following identity. This equality underlies the entire kernel trick.

Proposition 4.27. *For every $x, y \in X$,*

$$(\Phi_K(x), \Phi_K(y)) = K(x,y).$$

Proof. This is apparent: for every $x, y \in X$, we have

$$(\Phi_K(x), \Phi_K(y)) = \sum_{n=1}^{\infty} (\lambda_n^{1/2} u_n(x))(\lambda_n^{1/2} u_n(y)) = K(x,y).$$

□

5. Reproducing kernel Hilbert spaces

Our discussion in this section is informed heavily by [MNY06] and [CS02].

Let X be a locally compact, first countable topological space, μ a strictly positive, locally finite, σ-finite Borel measure on X, and K a continuous Mercer kernel with

$$\int_X K(x,x) dx < \infty.$$

Mercer theory furnishes an assignment

$$\{\text{Continuous Mercer kernels } K(x,y)\} \to \{\text{Continuous maps } \Phi : X \to L^2(X)\}$$
$$K \mapsto \Phi_K.$$

Note that the feature map Φ_K thus constructed is quite non-trivial - it relies on knowledge of the eigenfunctions $u_n(x)$ of the operator attached to a continuous Mercer kernel $K(x,y)$. Note too that, crucially, even talking about the feature map Φ_K relies on a serious fact - that any eigenfunction u of T_K with non-zero eigenvalue is continuous. This is because the Hilbert space $L^2(X)$ consists of *classes* of functions, and without something like continuity, there is no way to reasonably choose a representative of each class, and hence no way to make sense of *evaluating* an element of L^2 at a point.

This should motivate the following innocuous-seeming definition.

Definition 5.1. Let X be a set, H a vector subspace of the space of functions on X, equipped with an inner product $(\cdot,\cdot)$. We say that H is a *reproducing kernel Hilbert space* or *RKHS* (on X) if, for every $x \in X$, the linear functional

$$\begin{aligned} \mathrm{ev}_x : H &\to \mathbb{R} \\ \phi &\mapsto \phi(x) \end{aligned}$$

is continuous.

5.1. **Pointwise positive (semi-)definite kernels.** Let H be an RKHS on a set X. By the Riesz representation theorem (Proposition 2.11) there is, for every $x \in X$, an element $k_x \in H$ so that

$$\mathrm{ev}_x(\phi) = (k_x, \phi)$$

for all $\phi \in H$. We can consider the function

$$K(x,y) := \mathrm{ev}_y(k_x) = (k_y, k_x)$$

which is clearly a symmetric function of two variables.

The positive definite-ness of $(\cdot,\cdot)$ on H translates to the following property.

Definition 5.2. Let X be a set, $K : X \times X \to \mathbb{R}$ a function. K is *pointwise positive definite (p.p.d.)* if, for every finite list of points $x_1, \ldots, x_N \in X$, and constants $c_1, \ldots c_N \in \mathbb{R}$,

$$\sum_{m=1}^{N} \sum_{n=1}^{N} c_m c_n K(x_m, x_n) \geq 0$$

Lemma 5.3. *Let X be a set H an RKHS on X, and let $K(x,y) = (k_y, k_x)$ be the associated kernel function. Then K is pointwise positive definite.*

Proof. We simply compute

$$\sum_{m=1}^{N} \sum_{n=1}^{N} c_m c_n K(x_m, x_n) = \sum_{m=1}^{N} \sum_{n=1}^{N} c_m c_n (k_{x_n}, k_{x_m}) = (\sum_{n=1}^{N} c_n k_{x_n}, \sum_{m=1}^{N} c_m k_{x_m}) \geq 0$$

by positive definite-ness of the inner product on H. □

It is a remarkable observation that the notion of a RKHS is *equivalent* to that of a symmetric, p.p.d. kernel.

We need an easy fact from linear algebra.

Lemma 5.4. *Let V be a real vector space equipped with a positive semi-definite bilinear form $(\cdot,\cdot)$, i.e. a symmetric bilinear form satisfying*

$$(v,v) \geq 0$$

for all $v \in V$. Let $N = \mathrm{Rad}(V) := \{v \in V : (v,v) = 0\}$. Then N is a subspace of V, and for any $v \in V$ and $n \in N$, $(v,n) = 0$.

Proof. N is clearly closed under scalar multiplication. To see that it is closed under addition, let $n_1, n_2 \in N$. If $0 < (n_1+n_2, n_1+n_2)$ then, as $(n_1+n_2, n_1+n_2) = (n_1,n_1) + 2(n_1,n_2) + (n_2,n_2) = 2(n_1,n_2)$, $(n_1,n_2) > 0$. But then

$$(n_1 - n_2, n_1 - n_2) = -2(n_1, n_2) < 0$$

which violates positive semi-definiteness.

Similarly, if $v \in V, n \in N$, then if $(v,n) \neq 0$, we can compute

$$(v + \alpha n, v + \alpha n) = (v,v) + 2\alpha(v,n)$$

which can be made negative by choosing α appropriately, a contradiction. □

Remark 5.5. For V as above, and $N = \mathrm{Rad}(V)$, let V' be any vector-space complement to N. Then $(V', (\cdot,\cdot))$ is an inner product space, i.e. $(\cdot,\cdot)|_{V'\times V'}$ is strictly positive definite.

With this in hand, we proceed to the main result of this section.

Theorem 5.6 (Moore-Aronszajn)**.** *Let X be a set, $K : X \times X \to \mathbb{R}$ a symmetric, p.p.d. kernel. For each $x \in X$, set*

$$K_x = K(x, \cdot)$$

to be the "slice" of K (formally, we formally add a vector for each $x \in X$). Let $H_1 = \bigoplus_{x\in X} \mathbb{R}K_x$ be the algebraic span of the K_x, as x varies over X, and equip this vector space with the obvious "evaluation at x" functionals.

Define, for $\phi = \sum_{m=1}^{M} a_m K_{x_m}$, $\psi = \sum_{n=1}^{N} b_n K_{y_n}$,

$$(\phi, \psi) := \sum_{m=1}^{M}\sum_{n=1}^{N} a_m b_n K(x_m, y_n).$$

Then $(\cdot,\cdot)$ is a positive semi-definite symmetric bilinear form on H_0. Let H_0' be any vector space complement to $N_0 = \mathrm{Rad}(H_0)$, and let H' be the completion of H_0' with respect to $(\cdot,\cdot)$. Then H' is an RKHS on X with associated kernel K.

Proof. We need to check that

(1) The formula
$$(\phi, \psi) := \sum_{n=1}^{N}\sum_{m=1}^{M} a_i b_j K(x_i, y_j)$$
describes a positive semi-definite inner product on H_0.

(2) That, for given $x \in X$, evaluation at x is continuous on H_0'.

Neither of these is difficult. (1) follows immediately from the construction: given $\phi = \sum_{n=1}^{N} a_n K_{x_n} \in H_0$,

$$(\phi, \phi) := \sum_{m=1}^{N}\sum_{n=1}^{M} a_m a_n K(x_m, x_n) \geq 0$$

since K is pointwise positive definite.

As for (2), let $x_1, \dots x_N \in X$, $a_1, \dots a_N \in \mathbb{R}$, and suppose that $\phi = \sum_{n=1}^{N} a_n K_{x_n} \in H_0'$. Clearly, evaluation at x can be written as $\phi(x) = (K_x, \phi)$ - the issue is that K_x may not lie in H'. But, if we decompose K_x (uniquely) as

$$K_x = K_x' + N_x$$

where $K_x' \in H_0'$, $N_x \in N_0$, then

$$\phi(x) = (K_x, \phi) = (K_x', \phi) + (N_x, \phi) = (\phi, K_x')$$

so in fact N_x' also represents evaluation at x. □

Remark 5.7. It is somewhat remarkable that the notion of a p.p.d. kernel allows us to immediately identify a feature map $x \mapsto K_x$ to a Hilbert space, without any reference to spectral theory. [MNY06] emphasizes, though, that the spectrum of a Mercer kernel plays a central role in the error estimates for learning theory (for K) – the RKHS/Moore-Aronszajn approach to feature maps ignores this completely.

Definition 5.8. We denote by $\mathrm{MA}_K := H'$ the RKHS described in the Moore-Aronszajn theorem.

Let us now more carefully pin the down the relationship between RKHSs, p.p.d. kernels, and feature maps. First, let's be clear about the notion of a "feature map".

Definition 5.9. Let X be a set. A *feature map* (H, Φ) for X is the data of a Hilbert space H, together with a map $\Phi : X \to H$.

To every feature map (H, Φ), we may associate a symmetric kernel

$$K(x, y) = (\Phi(x), \Phi(y)).$$

By Lemma 5.3, this is a p.p.d. kernel.

Definition 5.10. We say a feature map (H, Φ) is "minimal" if there is no proper closed subspace $H' \subseteq$ containing the image of Φ.

Given a feature map (H, Φ), there is an obvious "minimalification." Let $H_{\Phi,0} := \mathrm{Span}(\{\Phi(x) : x \in X\})$, and $H_\Phi := \overline{H_{\Phi,0}}$ the closure of $H_{\Phi,0}$ in H. (H_Φ, Φ) is obviously a minimal feature map.

Given any minimal feature map (H', Φ), we can identify H' with a space of functions on X, by setting, for $\phi \in H'$,

$$\mathrm{ev}_x(\phi) = \phi(x) = (\phi, \Phi(x));$$

by the density of $H'_0 := \mathrm{Span}\{\Phi(x) : x \in X\}$ in H' this identifies H' with its image in the space of functions on X. It follows that H' is an RKHS with associated kernel K.

To summarize: let X be a set. We have the following constructions:

$$\begin{tikzcd}
\{\text{Feature maps } (H,\Phi)\} \arrow[r] \arrow[d] & \{\text{p.p.d. kernels } K\} \arrow[dl] \\
\{\text{Minimal feature maps } (H',\Phi)\} &
\end{tikzcd}$$

where the top arrow is $K(x, y) = (\Phi(x), \Phi(y)$; the downwards arrow is "minimalification" (adjoint to the inclusion); and the diagonal downwards left arrow sends K to MA_K.

There is essentially only one way to construct a minimal feature map (H', Φ) from K.

Lemma 5.11. *Let K be a p.p.d. kernel, and let (H'_1, Φ_1), (H'_2, Φ_2) be two minimal feature maps representing K. Then there is a unique isometry $\Psi : H'_1 \to H'_2$ so that*

$$\begin{tikzcd}
X \arrow[r, "\Phi_1"] \arrow[dr, "\Phi_2"'] & H'_1 \arrow[d, "\Psi"] \\
& H'_2
\end{tikzcd}$$

commutes.

Proof. We first define Ψ on $H'_{1,0} := \mathrm{Span}\{\Phi_1(x) : x \in X\}$ by simply sending

$$\Psi\Big(\sum_{n=1}^{N} a_n \Phi_1(x_n)\Big) = \sum_{n=1}^{N} a_n \Phi_2(x_n).$$

This is obviously linear - it is also an isometry, since

$$\begin{aligned}
\Big(\Psi\big(\sum_{n=1}^{N} a_n \Phi_1(x_n)\big), \Psi\big(\sum_{n=1}^{N} a_n \Phi_1(x_n)\big)\Big)_{H'_2} &= \Big(\sum_{n=1}^{N} a_n \Phi_2(x_n), \sum_{n=1}^{N} a_n \Phi_2(x_n)\Big)_{H'_2} \\
&= \sum_{m=1}^{N}\sum_{n=1}^{N} a_m a_n K(x_m, x_n) \\
&= \Big(\sum_{n=1}^{N} a_n \Phi_1(x_n), \sum_{n=1}^{N} a_n \Phi_1(x_n)\Big)_{H'_1}.
\end{aligned}$$

Ψ therefore extends by continuity to H'_1, landing in H'_2. The uniqueness of Ψ is also apparent. □

We now turn to providing some explicit examples of RKHSs and their associated kernels.

5.2. **Bandlimited functions: an example.** Let $X = \mathbb{R}^n$, and fix $\Xi \subseteq \mathbb{R}^n$ a compact rectangle with positive volume. Let $\mathscr{S}(\mathbb{R}^n)$ denote the space of Schwartz functions, i.e. functions ϕ which are smooth and have all partial derivatives rapidly decreasing. Let $H_0 \subseteq \mathscr{S}(\mathbb{R}^n)$ be the space of all ϕ whose Fourier transform

$$\mathscr{F}(\phi)(\xi) := \int_{\mathbb{R}^n} \phi(x) e^{-2\pi i x\cdot\xi} dx$$

is supported on Ξ; we equip H_0 with the L^2-inner product.

Remark 5.12. H_0 is sometimes called the space of band-limited Schwartz functions - it is the space of (inverse) Fourier transforms of smooth functions with support contained in Ξ.

Lemma 5.13. *Let $\phi \in H_0$. Then, for every $x \in \mathbb{R}^n$,*

$$|\phi(x)| \leq \|\phi\|_{L^2(\mathbb{R}^n)} \mathrm{vol}(\Xi)^{1/2}.$$

Proof. By Fourier inversion and Cauchy-Schwartz,

$$\begin{aligned} |\phi(x)| &= \left| \int_{\mathbb{R}^n} \mathscr{F}(\phi)(\xi) e^{2\pi i x \cdot \xi} \mathbb{1}_{\Xi}(\xi) d\xi \right| \\ &\leq \|\mathscr{F}(\phi)\|_{L^2(\mathbb{R}^n)} \mathrm{vol}(\Xi)^{1/2} \end{aligned}$$

but by Plancherel's theorem , $\|\mathscr{F}(\phi)\|_{L^2(\mathbb{R}^n)} = \|\phi\|_{L^2(\mathbb{R}^n)}$. ☐

Lemma 5.14. *Let H be the closure of the bandlimited Schwartz functions H_0 in L^2. Then H can be identified with the space of continuous L^2 functions with Fourier transform supported in Ξ.*

Proof. Let $\{\phi_n\}_{n=1}^{\infty}$ be a sequence of functions with $\phi_n \in H_0$ for all n, converging in $L^2(\mathbb{R}^n)$ to a function ϕ. We claim that ϕ_n then converges uniformly to ϕ - this is easy, since by Lemma 5.13

$$|\phi_n(x) - \phi_m(x)| \leq \|\phi_n - \phi_m\|_{L^2(\mathbb{R}^n)} \mathrm{vol}(\Xi)^{1/2}$$

so the sequence $\{\phi_n\}$ is uniformly Cauchy. It follows that the pointwise limit of the ϕ_n is continuous, as uniform limits of continuous functions are continuous.

It remains to see that every continuous L^2 function with Fourier transform supported on Ξ is the limit of functions in H_0. But this is easy - given $\phi \in L^2(\mathbb{R}^n) \cap C(\mathbb{R}^n)$ whose Fourier transform has support in Ξ, there exists, by density of the Schwartz space in L^2, a sequence $\phi_n \in \mathscr{S}(\mathbb{R}^n)$ with $\phi_n \to \phi$ in L^2. Now, $\phi_n \to \phi$ in L^2 if and only if $\mathscr{F}(\phi_n) \to \mathscr{F}(\phi)$ in L^2. We may fix compact rectangles $\Xi_1 \subset \Xi_2 \subset \ldots$ with $\Xi_i \subset \Xi$ and $\bigcup_i \Xi_i = \Xi^\circ$, and smooth functions j_i which are identically 1 on Ξ_i, have $0 \leq j_i(x) \leq 1$ for all x, and which are supported in Ξ. Passing to a subsequence i_k of indices if necessary, we find that $j_{i_k}\mathscr{F}(\phi_k) \to \mathscr{F}(\phi)$ as $k \to \infty$ in L^2. Taking inverse Fourier transforms allows us to conclude. ☐

The Hilbert space H of band-limited functions as described above is a reproducing kernel Hilbert space. Indeed, the functional "evaluation at x" is bounded by Lemma 5.13:

$$|\mathrm{ev}_x(g)| = |g(x)| \leq \|g\|_{L^2(\mathbb{R}^n)} \mathrm{vol}(\xi)^{1/2}.$$

If we explicitly write $\Xi = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$, then the representing function for evaluation at $x_0 = (x_1^{(0)}, \ldots, x_n^{(0)})$ can be pinned down completely.

Lemma 5.15.

$$k_{x_0}(x) = \frac{1}{\pi^n} \prod_{j=1}^{n} e^{\pi i (x_j - x_j^{(0)})(b_j + a_j)} \frac{\sin(\pi(x_j - x_j^{(0)})(b_j + a_j))}{x_j - x_j^{(0)}}$$

Proof. Let $x_0 = (x_1^{(0)}, \ldots, x_n^{(0)}) \in \mathbb{R}^n$, and k_{x_0} represent evaluation at x_0. By the Fourier inversion formula, we have, for $\phi \in H$,

$$\phi(x_0) = \int_{\mathbb{R}^n} \mathscr{F}(\phi)(\xi) e^{2\pi i x_0 \cdot \xi} d\xi = \int_{\mathbb{R}^n} \mathscr{F}(\phi)(\xi) e^{2\pi i x_0 \cdot \xi} \mathbb{1}_{\Xi}(\xi) d\xi$$

and by the Plancherel theorem, we can rewrite this as

$$\phi(x_0) = \int_{\mathbb{R}^n} \phi(x) \overline{\left(\int_{\Xi} e^{2\pi i (x - x_0) \cdot \xi} d\xi \right)} dx.$$

Computing, we find

$$\begin{aligned}
k_{x_0}(x) &= \int_\Xi e^{2\pi i(x-x_0)\cdot\xi} d\xi \\
&= \prod_{j=1}^n \frac{1}{2\pi i(x_j - x_j^{(0)})} \left(e^{2\pi i(x_j - x_j^0)b_j} - e^{2\pi i(x_j - x_j^0)a_j}\right) \\
&= \prod_{j=1}^n \frac{e^{\pi i(x_j - x_j^{(0)})(b_j+a_j)}}{2\pi i(x_j - x_j^{(0)})} \left(e^{\pi i(x_j - x_j^0)(b_j - a_j)} - e^{\pi i(x_j - x_j^0)(b_j - a_j)}\right) \\
&= \frac{1}{\pi^n} \prod_{j=1}^n e^{\pi i(x_j - x_j^{(0)})(b_j+a_j)} \frac{\sin(\pi(x_j - x_j^{(0)})(b_j + a_j))}{x_j - x_j^{(0)}}.
\end{aligned}$$

□

5.3. **The RKHS associated to a Mercer kernel.** Another class of examples comes from (the continuous form of) Mercer's theorem. Thus, throughout this section, X will be locally compact and first countable, and equipped with a Borel measure μ which is locally finite, totally positive, and σ-finite. K is a continuous Mercer kernel satisfying

$$\int_X K(x,x)dx < \infty.$$

First, we explain the relation between positive definiteness and pointwise positive definiteness.

Lemma 5.16. *Under the above assumptions, K is pointwise positive definite.*

Proof. Let $x_0 \in X$, and let $U_1 \supseteq U_2 \supseteq \dots$ be a neighborhood basis of x_0, with $0 < \mu(U_i) < \infty$. If we let $\delta_{U_i} = \frac{1}{\mu(U_i)} \mathbb{1}_{U_i}$, then

$$\lim_{i \to infty} (\delta_{U_i}, f) = f(x_0)$$

for every function f which is continuous on X.

So, given $c_1, \dots, c_N \in \mathbb{R}$ and $x_1, \cdot, x_N \in X$, let $U_1^{(n)} \supseteq U_2^{(n)} \supseteq \dots$ be a neighborhood basis of x_n for each n, as above. Set

$$\phi_i = \sum_{n=1}^N c_n \delta_{U_i^{(n)}}.$$

Observe that

$$\int_X \int_X \phi_i(x) K(x,y) \phi_i(y) dx dy \to \sum_{m=1}^N \sum_{n=1}^N c_m c_n K(x_m, x_n)$$

as $i \to \infty$. It follows that

$$\sum_{m=1}^N \sum_{n=1}^N c_m c_n K(x_m, x_n) \geq 0.$$

□

By the above, it makes sense to consider the RKHS MA_K. There is, however, a different-looking description of this RKHS.

Let $\{u_n : n = 1, 2, \dots\}$ be the orthonormal set of eigenfunctions of T_K with positive eigenvalues, ordered by eigenvalue.

Definition 5.17. Let $H_0 = \mathrm{Span}\{u_n : n = 1, 2, \dots\}$, and define on H_0 an inner product by setting,

$$(u_m, u_n)_H := \begin{cases} \lambda_n^{-1} & \text{if } m = n \\ 0 & \text{else} \end{cases}.$$

By construction, this is (strictly) positive definite on H_0. Let H be the Hilbert space completion of this inner product.

H comes equipped with an obvious orthonormal basis

$$\widetilde{u}_n := \lambda_n^{1/2} u_n,$$

so concretely, H can be thought of as the collection of formal sums

$$\phi = \sum_{n=1}^{\infty} \widetilde{a}_n \widetilde{u}_n : \widetilde{a}_n \in \mathbb{R}, \sum_{n=1}^{\infty} \widetilde{a}_n^2 < \infty\}.$$

If we write $a_n = \lambda_n^{1/2}\widetilde{a}_n$, then an element of H is a formal sum $\sum_{n=1}^{\infty} a_n u_n$ where $\sum_{n=1}^{\infty} \lambda_n^{-1} a_n^2 < \infty$. Since $\lambda_n \to 0$ as $n \to \infty$, it follows that if $\sum_{n=1}^{\infty} \lambda_n^{-1} a_n^2 < \infty$, then $\sum_{n=1}^{\infty} a_n^2 < \infty$ - thus $H \subseteq L^2(X)$.

Proposition 5.18. *Let $\phi = \sum_{n=1}^{\infty} a_n u_n$ be a vector in H. Then the series*

$$\sum_{n=1}^{\infty} a_n u_n(x)$$

converges absolutely and uniformly on compact sets of X. In particular, as an element of $L^2(X)$, ϕ has a representative which is a continuous function on X.

Proof. Indeed, by Cauchy-Schwartz and Proposition 4.12,

$$\begin{aligned} \left(\sum_{n=1}^{\infty} |a_n u_n(x)|\right)^2 &= \sum_{n=1}^{\infty} |\widetilde{a}_n \lambda_n^{1/2} u_n(x)|^2 \\ &\leq \left(\sum_{n=1}^{\infty} \widetilde{a}_n^2\right)\left(\sum_{n=1}^{\infty} \lambda_n u_n(x)^2\right) \\ &= \|\phi\|_H^2 K(x,x). \end{aligned}$$

□

Proposition 5.19. *Let $x_0 \in X$. The evaluation functional*

$$\mathrm{ev}_{x_0} : H \to \mathbb{R}$$

is continuous.

Proof. This follows almost immediately by the same argument:

$$\begin{aligned} |\mathrm{ev}_{x_0}(\phi)|^2 = |\phi(x_0)|^2 &= |\sum_{n=1}^{\infty} \widetilde{a}_n(\lambda_n^{1/2} u_n(x))| \\ &\leq \left(\sum_{n=1}^{\infty} \widetilde{a}_n^2\right)\left(\sum_{n=1}^{\infty} \lambda_n u_n(x)^2\right) = \|\phi\|_H^2 |K(x,x)|. \end{aligned}$$

□

It follows that H is an RKHS. It remains to compare it to the RKHS MA_K.

Proposition 5.20. MA_K *naturally isometrically embeds as a closed subspace of H.*

Proof. For this, we will proceed as follows:

(1) We will show that every slice $K_x = K(x, \cdot)$ of K lies in H.
(2) We will show that, for any $x, y \in X$,

$$(K_x, K_y)_H = K(x,y) = (K_x, K_y)_{\mathrm{MA}_K}.$$

Since the linear span of the slices is dense in MA_K, it follows that all of MA_K appears isometrically as a subspace of H.

It remains to show (1) and (2) above. For (1), we appeal to Mercer's theorem to expand the kernel function $K(x,y)$ as the sum, absolutely and uniformly convergent on compact sets,

$$K(x,y) = \sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y).$$

We may write the slices of K as

$$K_x = K(x,\cdot) = \sum_{n=1}^{\infty} \lambda_n^{1/2} u_n(x) \widetilde{u}_n(\cdot).$$

Since, by assumption, $K(x,x) < \infty$, we have that

$$\sum_{n=1}^{\infty} \lambda_n u_n(x)^2 < \infty$$

for every x - in particular,

$$(\lambda_n^{1/2} u_n(x))_{n=1}^{\infty} \in l^2(\mathbb{N})$$

for each x, i.e.

$$K(x,\cdot) = \sum_{n=1}^{\infty} \lambda_n^{1/2} u_n(x) \widetilde{u}_n(\cdot)$$

is an element of H.

For (2), we simply compute that for $K_x = K(x,\cdot), K_y = K(y,\cdot)$, the pairing

$$(K_x, K_y)_H = ((\lambda_n^{1/2} u_n(x))_{n=1}^{\infty}, (\lambda_n^{1/2} u_n(y))_{n=1}^{\infty})_{l^2(\mathbb{N})} = \sum_{n=1}^{\infty} \lambda_n u_n(x) u_n(y) = K(x,y) = (K_x, K_y)_{\mathrm{MA}_K}.$$

$\square$

Remark 5.21. Note that the feature map here is very different than the one occurring in our discussion of Mercer's theorem - it lands in an RKHS rather than a space of classes of functions on X (i.e. $L^2(X)$).

References

[CS02] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[MHC12] A. P. Peron M. H. Castro, V. A. Menegatto. Integral operators generated by mercer-like kernels on topological spaces. *Colloquium Mathematicae*, 126(1):125–138, 2012.

[MNY06] Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In Gábor Lugosi and Hans Ulrich Simon, editors, *Learning Theory*, pages 154–168, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.