Introduction to RKHS, and some simple kernel algorithms

Arthur Gretton

October 16, 2019

1 Outline

In this document, we give a nontechical introduction to reproducing kernel Hilbert spaces (RKHSs), and describe some basic algorithms in RHKS.

1. What is a kernel, how do we construct it?
2. Operations on kernels that allow us to combine them
3. The reproducing kernel Hilbert space
4. Application 1: Difference in means in feature space
5. Application 2: Kernel PCA
6. Application 3: Ridge regression

2 Motivating examples

For the XOR example, we have variables in two dimensions, $x \in \mathbb{R}^2$, arranged in an XOR pattern. We would like to separate the red patterns from the blue, using only a linear classifier. This is clearly not possible, in the original space. If we map points to a higher dimensional feature space

$$\phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3,$$

it is possible to use a linear classifier to separate the points. See Figure 2.1.

Feature spaces can be used to compare objects which have much more complex structure. An illustration is in Figure 2.2, where we have two sets of documents (the red ones on dogs, and the blue on cats) which we wish to classify. In this case, features of the documents are chosen to be histograms over words (there are much more sophisticated features we could use, eg string kernels [4]). To use the terminology from the first example, these histograms represent a mapping of the documents to feature space. Once we have histograms, we can compare documents, classify them, cluster them, etc.

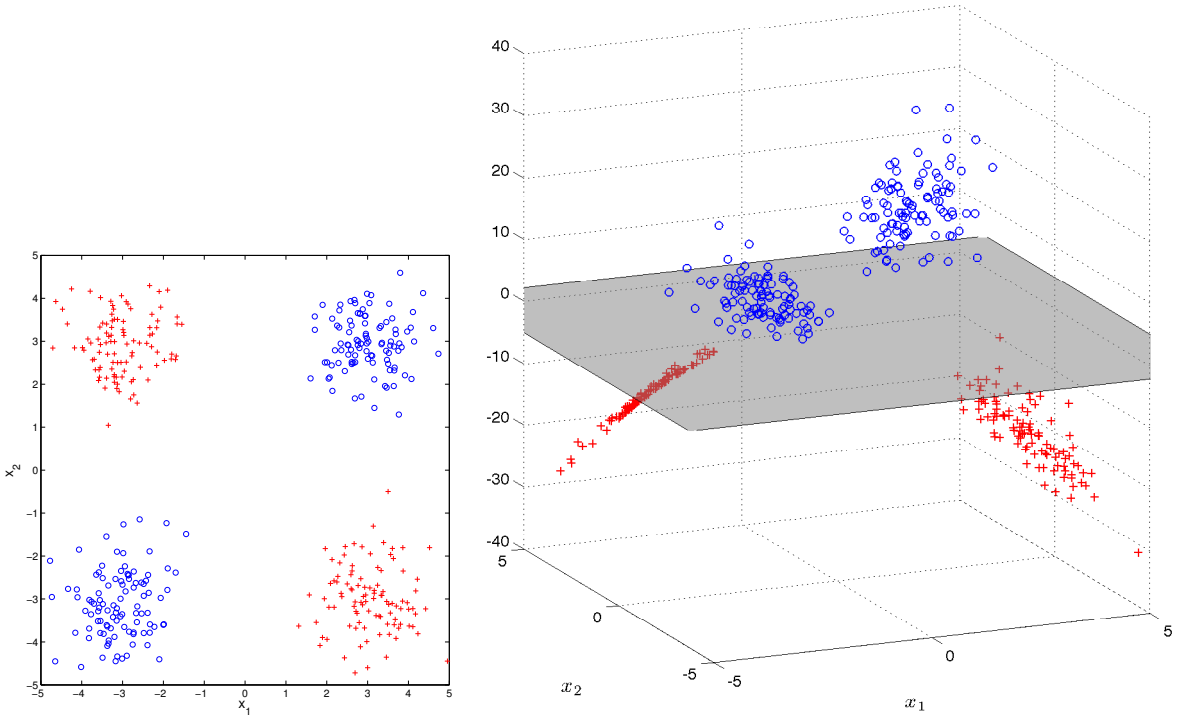

Figure 2.1: XOR example. On the left, the points are plotted in the original space. There is no linear classifier that can separate the red crosses from the blue circles. Mapping the points to a higher dimensional feature space, we obtain linearly separable classes. A possible decision boundary is shown as a gray plane.

The classsification of objects via well chosen features is of course not an unusual approach. What distinguishes kernel methods is that they can (and often do) use *infinitely many features*. This can be achieved as long as our learning algorithms are defined in terms of *dot products* between the features, where these dot products can be computed in closed form. The term "kernel" simply refers to a dot product between (possibly infinitely many) features.

Alternatively, kernel methods can be used to control smoothness of a function used in regression or classification. An example is given in Figure 2.3, where different parameter choices determine whether the regression function overfits, underfits, or fits optimally. The connection between feature spaces and smoothness is not obvious, and is one of the things we'll discuss in the course.

3 What is a kernel and how do we construct it?

3.1 Construction of kernels

The following is taken mainly from [11, Section 4.1].

Definition 1 (Inner product)**.** Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be *an inner product* on $\mathcal{H}$ if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$ [1]
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$.

We can define a norm using the inner product as $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

A Hilbert space is a space on which an inner product is defined, along with an additional technical condition.[2] We now define a kernel.

Definition 3. Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} .$$

Note that we imposed almost no conditions on $\mathcal{X}$: we don't even require there to be an inner product defined on the elements of $\mathcal{X}$. The case of documents is an instructive example: you can't take an inner product between two books, but you can take an inner product between features of the text.

[1] If the inner product is complex valued, we have conjugate symmetry, $\langle f, g \rangle_{\mathcal{H}} = \overline{\langle g, f \rangle}_{\mathcal{H}}$.

[2] Specifically, a Hilbert space must contain the limits of all Cauchy sequences of functions:

Definition 2 (Cauchy sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements in a normed space $\mathcal{H}$ is said to be *a Cauchy sequence* if for every $\epsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$, such that for all $n, m \geq N$, $\|f_n - f_m\|_{\mathcal{H}} < \epsilon$.

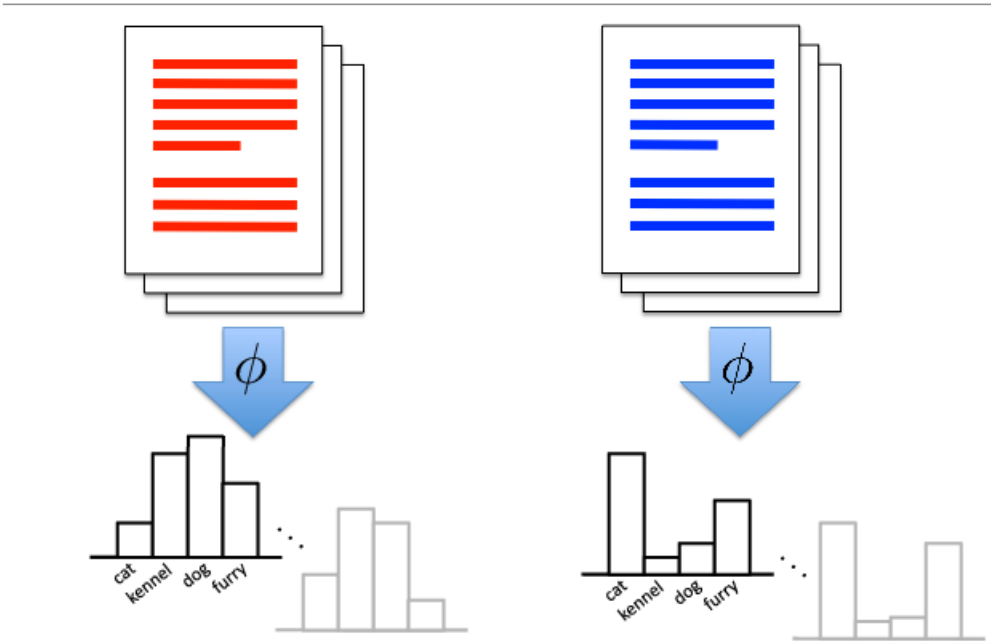

Figure 2.2: Document classification example: each document is represented as a histogram over words.

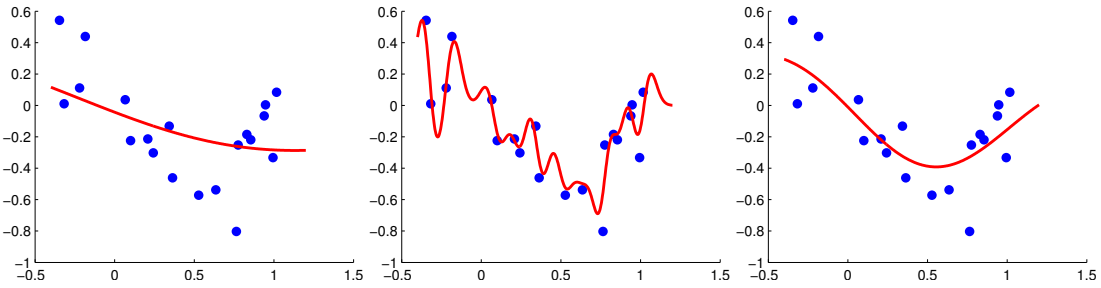

Figure 2.3: Regression: examples are shown of underfitting, overfitting, and a good fit. Kernel methods allow us to control the smoothness of the regression function.

A single kernel can correspond to multiple sets of underlying features. A trivial example for $\mathcal{X} := \mathbb{R}$:

$$\phi_1(x) = x \qquad \text{and} \qquad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$

Kernels can be combined and modified to get new kernels:

Lemma 4 (Sums of kernels are kernels)**.** *Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.*

To easily prove the above, we will need to use a property of kernels introduced later, namely *positive definiteness*. We provide this proof at the end of Section 3.2. A difference of kernels may not be a kernel: if $k_1(x, x) - k_2(x, x) < 0$, then condition 3 of Definition 1 is violated.

Lemma 5 (Mappings between spaces)**.** *Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \widetilde{\mathcal{X}}$. Define the kernel k on $\widetilde{\mathcal{X}}$. Then the kernel $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.*

Lemma 6 (Products of kernels are kernels)**.** *Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $k := k_1 \times k_2$ is a kernel on $\mathcal{X}$.*

Proof. The general proof has some technicalities: see [11, Lemma 4.6 p. 114]. However, the main idea can be shown with some simple linear algebra. We consider the case where $\mathcal{H}_1$ corresponding to k_1 is $\mathbb{R}^m$, and $\mathcal{H}_2$ corresponding to k_2 is $\mathbb{R}^n$. Define $k_1 := u^\top v$ for vectors $u, v \in \mathbb{R}^m$ and $k_2 := p^\top q$ for vectors $p, q \in \mathbb{R}^n$.

We will use that the inner product between matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$ is

$$\langle A, B \rangle = \operatorname{trace}(A^\top B). \tag{3.1}$$

Then

$$\begin{aligned} k_1 k_2 &= k_1\left(p^\top q\right) \\ & \quad k_1\left(q^\top p\right) \\ &= k_1 \mathrm{trace}(q^\top p) \\ &= k_1 \mathrm{trace}(pq^\top) \\ &= \mathrm{trace}(p\underbrace{u^\top v}_{k_1} q^\top) \\ &= \langle A, B\rangle, \end{aligned}$$

where we defined $A := up^\top$ and $B := vq^\top$. In other words, the product $k_1 k_2$ defines a valid inner product in accordance with (3.1). □

The sum and product rules allow us to define a huge variety of kernels.

Lemma 7 (Polynomial kernels)**.** *Let $x, x' \in \mathbb{R}^d$ for $d \geq 1$, and let $m \geq 1$ be an integer and $c \geq 0$ be a positive real. Then*

$$k(x, x') := \left(\langle x, x'\rangle + c\right)^m$$

is a valid kernel.

To prove: expand out this expression into a sum (with non-negative scalars) of kernels $\langle x, x'\rangle$ raised to integer powers. These individual terms are valid kernels by the product rule.

Can we extend this combination of sum and product rule to sums with infinitely many terms? It turns out we can, as long as these don't blow up.

Definition 8. The space ℓ_p of p-summable sequences is defined as all sequences $(a_i)_{i\geq 1}$ for which

$$\sum_{i=1}^{\infty} a_i^p < \infty.$$

Kernels can be defined in terms of sequences in ℓ_2.

Lemma 9. *Given a non-empty set $\mathcal{X}$, and a sequence of functions $(\phi_i(x))_{i\geq 1}$ in ℓ_2 where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the ith coordinate of the feature map $\phi(x)$. Then*

$$k(x, x') := \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x') \tag{3.2}$$

is a well-defined kernel on $\mathcal{X}$.

Proof. We write the norm $\|a\|_{\ell_2}$ associated with the inner product (3.2) as

$$\|a\|_{\ell_2} := \sqrt{\sum_{i=1}^{\infty} a_i^2},$$

where we write a to represent the sequence with terms a_i. The Cauchy-Schwarz inequality states

$$|k(x,x')| = \left|\sum_{i=1}^{\infty} \phi_i(x)\phi_i(x')\right| \leq \|\phi(x)\|_{\ell_2} \|\phi(x')\|_{\ell_2} < \infty,$$

so the kernel in (3.2) is well defined for all $x, x' \in \mathcal{X}$. □

Taylor series expansions may be used to define kernels that have infinitely many features.

Definition 10. [Taylor series kernel] [11, Lemma 4.8] Assume we can define the Taylor series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \qquad |z| < r,\ z \in \mathbb{R},$$

for $r \in (0, \infty]$, with $a_n \geq 0$ for all $n \geq 0$. Define $\mathcal{X}$ to be the $\sqrt{r}$-ball in $\mathbb{R}^d$. Then for $x, x' \in \mathbb{R}^d$ such that $\|x\| < \sqrt{r}$, we have the kernel

$$k(x,x') = f\left(\langle x, x'\rangle\right) = \sum_{n=0}^{\infty} a_n \langle x, x'\rangle^n.$$

Proof. Non-negative weighted sums of kernels are kernels, and products of kernels are kernels, so the following is a kernel **if it converges,**

$$k(x,x') \quad = \quad \sum_{n=0}^{\infty} a_n \left(\langle x, x'\rangle\right)^n.$$

We have by Cauchy-Schwarz that

$$|\langle x, x'\rangle| \leq \|x\|\|x'\| < r,$$

so the Taylor series converges. □

An example of a Taylor series kernel is the exponential.

Example 11 (Exponential kernel)**.** The exponential kernel on $\mathbb{R}^d$ is defined as

$$k(x,x') := \exp\left(\langle x, x'\rangle\right).$$

We may combine all the results above to obtain the following (the proof is an exercise - you will need the product rule, the mapping rule, and the result of Example 11).

Example 12 (Gaussian kernel)**.** The Gaussian kernel on $\mathbb{R}^d$ is defined as

$$k(x,x') := \exp\left(-\gamma^{-2}\|x - x'\|^2\right).$$

3.2 Positive definiteness of an inner product in a Hilbert space

All kernel functions are **positive definite**. In fact, if we have a positive definite function, we know there exists one (or more) feature spaces for which the kernel defines the inner product - we are not obliged to define the feature spaces explicitly. We begin by defining positive definiteness [1, Definition 2], [11, Definition 4.15].

Definition 13 (Positive definite functions)**.** A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, \dots a_n) \in \mathbb{R}^n$, $\forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

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

The function $k(\cdot, \cdot)$ is *strictly* positive definite if for mutually distinct x_i, the equality holds only when all the a_i are zero.[3]

Every inner product is a positive definite function, and more generally:

Lemma 14. *Let $\mathcal{H}$ be any Hilbert space (not necessarily an RKHS), $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function.*

Proof.

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{H}} . \\ & \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0 \end{aligned}$$

□

Remarkably, the reverse direction also holds: a positive definite function is guaranteed to be the inner product in a Hilbert space $\mathcal{H}$ between features $\phi(x)$ (which we need not even specify explicitly). The proof is not difficult, but has some technical aspects: see [11, Theorem 4.16 p. 118].

Positive definiteness is the easiest way to prove a sum of kernels is a kernel. Consider two kernels $k_1(x, x')$ and $k_2(x, x')$. Then

[3] Wendland [12, Definition 6.1 p. 65] uses the terminology "positive semi-definite" vs "positive definite".

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left[k_1(x_i, x_j) + k_2(x_i, x_j)\right]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_2(x_i, x_j)$$
$$\geq 0$$

4 The reproducing kernel Hilbert space

We have introduced the notation of feature spaces, and kernels on these feature spaces. What's more, we've determined that these kernels are positive definite. In this section, we use these kernels to define *functions* on $\mathcal{X}$. The space of such functions is known as a reproducing kernel Hilbert space.

4.1 Motivating examples

4.1.1 Finite dimensional setting

We start with a simple example using the same finite dimensional feature space we used in the XOR motivating example (Figure 2.1). Consider the feature map

$$\begin{array}{rcl} \phi : \mathbb{R}^2 & \to & \mathbb{R}^3 \\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} & \mapsto & \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}, \end{array}$$

with the kernel

$$k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix}$$

(i.e., the standard inner product in $\mathbb{R}^3$ between the features). This is a valid kernel in the sense we've considered up till now. We denote by $\mathcal{H}$ this feature space.

Let's now define a function of the features x_1, x_2, and $x_1 x_2$ of x, namely:

$$f(x) = a x_1 + b x_2 + c x_1 x_2.$$

This function is a member of a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to $\mathbb{R}$. We can define an equivalent representation for f,

$$f(\cdot) = \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{4.1}$$

The notation $f(\cdot)$ refers to the function itself, in the abstract (and in fact, this function might have multiple equivalent representations). We sometimes write f rather than $f(\cdot)$, when there is no ambiguity. The notation $f(x) \in \mathbb{R}$ refers to the function evaluated at a particular point (which is just a real number). With this notation, we can write

$$\begin{aligned} f(x) &= f(\cdot)^\top \phi(x) \\ &:= \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} \end{aligned}$$

In other words, the evaluation of f at x can be written as an **inner product in feature space** (in this case, just the standard inner product in $\mathbb{R}^3$), and $\mathcal{H}$ is a space of functions mapping $\mathbb{R}^2$ to $\mathbb{R}$. This construction can still be used if there are infinitely many features: from the Cauchy-Schwarz argument in Lemma 9, we may write

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \phi_\ell(x), \tag{4.2}$$

where the expression is bounded in absolute value as long as $\{f_\ell\}_{\ell=1}^{\infty} \in \ell_2$ (of course, we can't write this function explicitly, since we'd need to enumerate all the f_ℓ).

This line of reasoning leads us to a conclusion that might seem counterintuitive at first: we've seen that $\phi(x)$ is a mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$, but it also defines (the parameters of) a *function* mapping $\mathbb{R}^2$ to $\mathbb{R}$. To see why this is so, we write

$$k(\cdot, y) = \begin{bmatrix} y_1 \\ y_2 \\ y_1 y_2 \end{bmatrix} = \phi(y),$$

using the same convention as in (4.1). This is certainly valid: if you give me a y, I'll give you a vector $k(\cdot, y)$ in $\mathcal{H}$ such that

$$\langle k(\cdot, y), \phi(x) \rangle_{\mathcal{H}} = ax_1 + bx_2 + cx_1x_2,$$

where $a = y_1$, $b = y_2$, and $c = y_1 y_2$ (i.e., for every y, we get a different vector $\begin{bmatrix} a & b & c \end{bmatrix}^\top$). But due to the symmetry of the arguments, we could equally well have written

$$\begin{aligned} \langle k(\cdot, x), \phi(y) \rangle &= uy_1 + vy_2 + wy_1y_2 \\ &= k(x, y). \end{aligned}$$

In other words, we can write $\phi(x) = k(\cdot, x)$ and $\phi(y) = k(\cdot, y)$ without ambiguity. This way of writing the feature mapping is called the **canonical feature map** [11, Lemma 4.19].

This example illustrates the two defining features of an RKHS:

- The feature map of every point is in the feature space:

$$\forall x \in \mathcal{X}, \;\; k(\cdot, x) \in \mathcal{H}$$

,

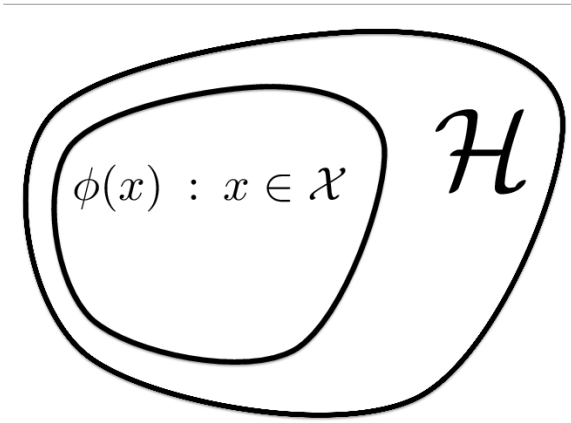

Figure 4.1: Feature space and mapping of input points.

- **The reproducing property**:

$$\forall x \in \mathcal{X},\, \forall f \in \mathcal{H},\ \ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x) \tag{4.3}$$

In particular, for any $x, y \in \mathcal{X}$,

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

Another, more subtle point to take from this example is that $\mathcal{H}$ is in this case larger than the set of all $\phi(x)$ (see Figure 4.1). For instance, when writing f in (4.1), we could choose $f = [1\ 1\ -1] \in \mathcal{H}$, but this cannot be obtained by the feature map $\phi(x) = [x_1\ x_2\ (x_1 x_2)]$.

4.1.2 Example: RKHS defined by via a Fourier series

Consider a function on the interval $[-\pi, \pi]$ with periodic boundary conditions. This may be expressed as a Fourier series,

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(\imath l x),$$

using the orthonormal basis on $[-\pi, \pi]$, noting that

$$\frac{1}{2}\int_{-\pi}^{\pi} \exp(\imath \ell x)\overline{\exp(\imath m x)}dx = \begin{cases} 1 & \ell = m, \\ 0 & \ell \neq m, \end{cases}$$

where $\imath = \sqrt{-1}$, and $\bar{a}$ is the complex conjugate of a. We assume $f(x)$ is real, so its Fourier transform is conjugate symmetric,

$$\hat{f}_{-\ell} = \overline{\hat{f}_\ell}.$$

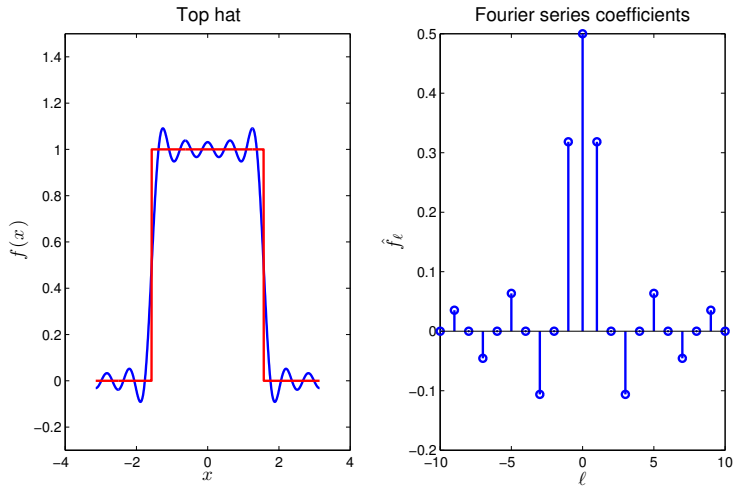

Figure 4.2: "Top hat" function (red) and its approximation via a Fourier series (blue). Only the first 21 terms are used; as more terms are added, the Fourier representation gets closer to the desired function.

As an illustration, consider the "top hat" function,

$$f(x) = \begin{cases} 1 & |x| < T, \\ 0 & T \leq |x| < \pi, \end{cases}$$

with Fourier series

$$\hat{f}_\ell := \frac{\sin(\ell T)}{\ell \pi} \qquad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_\ell \cos(\ell x).$$

Due to the symmetry of the Fourier coefficients and the asymmetry of the sine function, the sum can be written over positive ℓ, and only the cosine terms remain. See Figure 4.2.
Assume the kernel takes a single argument, which is the difference in its inputs,

$$k(x, y) = k(x - y),$$

and define the Fourier series representation of k as

$$k(x) = \sum_{l=-\infty}^{\infty} \hat{k}_l \exp(\imath l x), \tag{4.4}$$

where $\hat{k}_{-l} = \hat{k}_l$ and $\overline{\hat{k}_l} = \hat{k}_l$ (a real and symmetric $k$has a real and symmetric Fourier transform). For instance, when the kernel is a Jacobi Theta function ϑ (which looks close to a Gaussian when σ^2 is sufficiently narrower than $[-\pi, \pi]$),

$$k(x) = \frac{1}{2\pi}\vartheta\left(\frac{x}{2\pi}, \frac{\imath\sigma^2}{2\pi}\right), \qquad \hat{k}_\ell \approx \frac{1}{2\pi}\exp\left(\frac{-\sigma^2\ell^2}{2}\right),$$

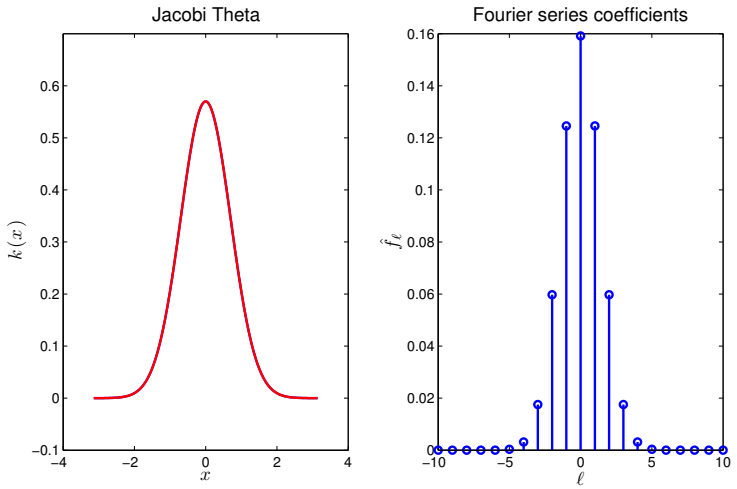

Figure 4.3: Jacobi Theta kernel (red) and its Fourier series representation, which is Gaussian (blue). Again, only the first 21 terms are retained, however the approximation is already very accurate (bearing in mind the Fourier series coefficients decay exponentially).

and the Fourier coefficients are Gaussian (evaluated on a discrete grid). See Figure 4.3.

Recall the standard dot product in L_2, where we take the conjugate of the right-hand term due to the complex valued arguments,

$$\begin{aligned}\langle f, g\rangle_{L_2} &= \left\langle \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \sum_{m=-\infty}^{\infty} \overline{\hat{g}_m \exp(\imath m x)} \right\rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell \left\langle \exp(\imath \ell x), \exp(-\imath m x)\right\rangle_{L_2} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \overline{\hat{g}}_\ell.\end{aligned}$$

We define the dot product in $\mathcal{H}$ to be a roughness penalized dot product, taking the form[4]

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{g}}_\ell}{\hat{k}_\ell}. \tag{4.5}$$

[4]Note: while this dot product has been defined using the Fourier transform of the kernel, additional technical conditions are required of the kernel for a valid RKHS to be obtained. These conditions are given by Mercer's theorem [11, Theorem 4.49], which when satisfied, imply that the expansion (4.4) converges absolutely and uniformly.

The squared norm of a function f in $\mathcal{H}$ enforces smoothness:

$$\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{f}_\ell}}{\hat{k}_\ell} = \sum_{l=-\infty}^{\infty} \frac{\left|\hat{f}_\ell\right|^2}{\hat{k}_\ell}. \tag{4.6}$$

If $\hat{k}_\ell$ decays fast, then so must $\hat{f}_\ell$ if we want $\|f\|_{\mathcal{H}}^2 < \infty$. From this norm definition, we see that the RKHS functions are a subset of the functions in L_2, for which finiteness of the norm $\|f\|_{L_2}^2 = \sum_{\ell=-\infty}^{\infty} \left|\hat{f}_\ell\right|^2$ is required (this being less restrictive than 4.6).

We next check whether the reproducing property holds for a function $f(x) \in \mathcal{H}$. Define a function

$$g(x) := k(x-z) = \sum_{\ell=-\infty}^{\infty} \exp(\imath\ell x) \underbrace{\hat{k}_\ell \exp(-\imath\ell z)}_{\hat{g}_\ell}$$

Then for a function[5] $f(\cdot) \in \mathcal{H}$,

$$\begin{aligned} \langle f(\cdot), k(\cdot, z)\rangle_{\mathcal{H}} &= \langle f(\cdot), g(\cdot)\rangle_{\mathcal{H}} \\ &= \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \left(\overline{\hat{k}_\ell \exp(-\imath\ell z)}\right)}{\hat{k}_\ell} \\ &= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath\ell z) \;=\; f(z). \end{aligned} \tag{4.7}$$

Finally, as a special case of the above, we verify the reproducing property for the kernel itself. Recall kernel definition,

$$k(x-y) = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath\ell(x-y)) = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath\ell x)\exp(-\imath\ell y)$$

Define two functions of a variable x as kernels centered at y and z, respectively,

$$\begin{aligned} f(x) := k(x-y) &= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(\imath\ell(x-y)) \\ &= \sum_{\ell=-\infty}^{\infty} \exp(\imath\ell x) \underbrace{\hat{k}_\ell \exp(-\imath\ell y)}_{\hat{f}_\ell} \\ g(x) := k(x-z) &= \sum_{\ell=-\infty}^{\infty} \exp(\imath\ell x) \underbrace{\hat{k}_\ell \exp(-\imath\ell z)}_{\hat{g}_\ell} \end{aligned}$$

[5]Exercise: what happens if we change the order, and write $\langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}}$? Hint: $f(x) = \overline{f(x)}$ since the function is real-valued.

Applying the dot product definition in $\mathcal{H}$, we obtain

$$\begin{aligned}\langle k(\cdot,y),k(\cdot,z)\rangle_{\mathcal{H}} &= \langle f,g\rangle_{\mathcal{H}}\\ &= \sum_{\ell=-\infty}^{\infty}\frac{\hat{f}_\ell\overline{\hat{g}_\ell}}{\hat{k}_\ell}\\ &= \sum_{\ell=-\infty}^{\infty}\frac{\left(\hat{k}_\ell\exp(-\imath\ell y)\right)\left(\overline{\hat{k}_\ell\exp(-\imath\ell z)}\right)}{\hat{k}_\ell}\\ &= \sum_{\ell=-\infty}^{\infty}\hat{k}_\ell\exp(\imath\ell(z-y)) \;=\; k(z-y).\end{aligned}$$

You might be wondering how the dot product in (4.5) relates to our original definition of an RKHS function in (4.2): the latter equation, updated to reflect that the features are complex-valued (and changing the sum index to run from $-\infty$ to ∞) is

$$f(x)=\sum_{\ell=-\infty}^{\infty}f_\ell\overline{\phi_\ell(x)},$$

which is an expansion in terms of coefficients f_ℓ and features $\phi_\ell(x)$. Writing the dot product from (4.7) earlier,

$$\begin{aligned}\langle f(\cdot),k(\cdot,z)\rangle_{\mathcal{H}} &= \sum_{\ell=-\infty}^{\infty}\frac{\hat{f}_\ell\left(\overline{\hat{k}_\ell\exp(-\imath\ell z)}\right)}{\hat{k}_\ell}\\ &= \sum_{\ell=-\infty}^{\infty}\frac{\hat{f}_\ell}{\sqrt{\hat{k}_\ell}}\left(\sqrt{\hat{k}_\ell}\left(\overline{\exp(-\imath\ell z)}\right)\right),\end{aligned}$$

it's clear that

$$f_\ell=\frac{\hat{f}_\ell}{\sqrt{\hat{k}_\ell}},\qquad\qquad \phi_\ell(x)=\sqrt{\hat{k}_\ell}\left(\exp(-\imath\ell z)\right),$$

and for this feature definition, the reproducing property holds,

$$\begin{aligned}\langle\phi_\ell(x),\phi_\ell(x')\rangle_{\mathcal{H}} &= \sum_{\ell=-\infty}^{\infty}\phi_\ell(x)\overline{\phi_\ell(x')}\\ &= \sum_{\ell=-\infty}^{\infty}\left(\sqrt{\hat{k}_\ell}\left(\exp(-\imath\ell x)\right)\right)\left(\sqrt{\hat{k}_\ell}\left(\overline{\exp(-\imath\ell x')}\right)\right)\\ &= k(x'-x).\end{aligned}$$

4.1.3 Example: RKHS defined using the exponentiated quadratic kernel on $\mathbb{R}$

Let's now consider the more general setting of kernels on $\mathbb{R}$, where we can no longer use the Fourier series expansion (the arguments in this section also apply

to the multivariate case $\mathbb{R}^d$). Our discussion follows [2, Sections 3.1 - 3.3]. We start by defining the eigenexpansion of $k(x, x')$ with respect to a non-negative finite measure μ on $\mathcal{X} := \mathbb{R}$,

$$\lambda_i e_i(x) = \int k(x, x') e_i(x') d\mu(x'), \qquad \int_{L_2(\mu)} e_i(x) e_j(x) d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{4.8}$$

For the purposes of this example, we'll use the Gaussian density μ, meaning

$$d\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2\right) dx \tag{4.9}$$

We can write

$$k(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x'), \tag{4.10}$$

which converges in $L_2(\mu)$.[6] If we choose an exponentiated quadratic kernel,

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right),$$

the eigenexpansion is

$$\begin{aligned} \lambda_k &\propto b^k \qquad b < 1 \\ e_k(x) &\propto \exp(-(c - a)x^2) H_k(x\sqrt{2c}), \end{aligned}$$

where a, b, c are functions of σ, and H_k is kth order Hermite polynomial [6, Section 4.3]. Three eigenfunctions are plotted in Figure 4.4.

We are given two functions f, g in $L_2(\mu)$, expanded in the orthonormal system $\{e_\ell\}_{\ell=1}^{\infty}$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x) \qquad g(x) = \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x), \tag{4.11}$$

The standard dot product in $L_2(\mu)$ between f, g is

$$\begin{aligned} \langle f, g \rangle_{L_2(\mu)} &= \left\langle \sum_{\ell=1}^{\infty} \hat{f}_\ell e_\ell(x), \sum_{\ell=1}^{\infty} \hat{g}_\ell e_\ell(x) \right\rangle_{L_2(\mu)} \\ &= \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell. \end{aligned}$$

As with the Fourier case, we will define the dot product in $\mathcal{H}$ to have a roughness penalty, yielding

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \qquad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell}, \tag{4.12}$$

[6] As with the Fourier example in Section (4.1.2), there are certain technical conditions needed when defining an RKHS kernel, to ensure that the sum in (4.10) converges in a stronger sense than $L_2(\mu)$. This requires a generalization of Mercer's theorem to non-compact domains.

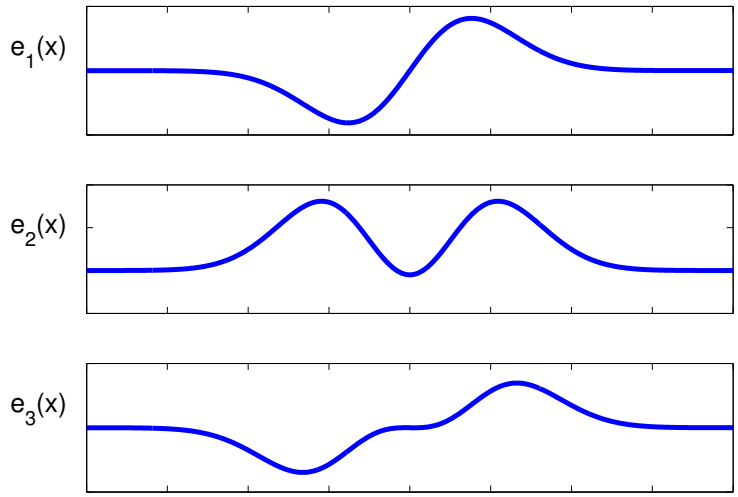

Figure 4.4: First three eigenfunctions for the exponentiated quadratic kernel with respect to a Gaussian measure μ.

where you should compare with (4.5) and (4.6). The RKHS functions are a subset of the functions in $L_2(\mu)$, with norm $\|f\|_{L_2}^2 = \sum_{\ell=1}^{\infty} \hat{f}_\ell^2 < \infty$ (less restrictive than 4.12).

Also just like the Fourier case, we can explicitly construct the feature map that gives our original expression of the RHKS function in (4.2), namely

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \phi_\ell(x).$$

We write the kernel centered at x as

$$g(x) := k(x - z) = \sum_{\ell=1}^{\infty} e_\ell(x) \underbrace{\lambda_\ell e_\ell(z)}_{\hat{g}_\ell}$$

Beginning with (4.11), we get

$$\begin{aligned} f(x) = \langle f, g \rangle_{\mathcal{H}} &= \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \\ &= \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \left(\lambda_\ell e_\ell(z)\right)}{\lambda_\ell} \\ &= \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell}{\sqrt{\lambda_\ell}} \left(\sqrt{\lambda_\ell} e_\ell(z)\right), \end{aligned}$$

hence

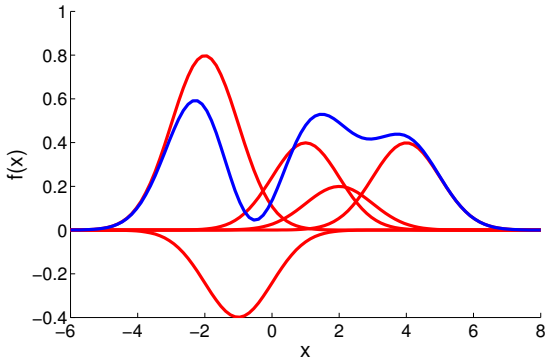

Figure 4.5: An RKHS function. The kernel here is an exponentiated quadratic. The blue function is obtained by taking the sum of red kernels, which are centred at x_i and scaled by α_i.

$$f_\ell = \frac{\hat{f}_\ell}{\sqrt{\lambda_\ell}} \qquad \phi_\ell(x) = \sqrt{\lambda_\ell} e_\ell(x), \tag{4.13}$$

and the reproducing property holds,[7]

$$\sum_{\ell=1}^{\infty} \phi_\ell(x)\phi_\ell(x') = k(x, x').$$

4.1.4 Tractable form of functions in an infinite dimensional RKHS, and explicit feature space construction

When a feature space is infinite dimensional, functions are generally expressed as linear combinations of kernels at particular points, such that the features need never be explicitly written down. The key is to satisfy the reproducing property in eq. (4.3) (and in Definition (15) below). Let's see, as an example, an RKHS function for an exponentiated quadratic kernel,

$$f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x). \tag{4.14}$$

We show an example function in Figure 4.5.

[7] Note also that the features are square summable, since

$$\begin{aligned} \|\phi(x)\|_{\mathcal{H}}^2 = \|\phi(x)\|_{\ell_2}^2 &= \sum_{\ell=1}^{\infty} \lambda_\ell e_\ell(x) e_\ell(x) \\ &= k(x, x) < \infty. \end{aligned}$$

The eigendecomposition in (4.10) and the feature definition in (4.13) yield

$$k(x,x') = \sum_{\ell=1}^{\infty} \underbrace{\left[\sqrt{\lambda_\ell} e_\ell(x)\right]}_{\phi_\ell(x)} \underbrace{\left[\sqrt{\lambda_\ell} e_\ell(x')\right]}_{\phi_\ell(x')}.$$

and (4.14) can be rewritten

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \phi_\ell(x) = \sum_{\ell=1}^{\infty} f_\ell \underbrace{\left[\sqrt{\lambda_\ell} e_\ell(x)\right]}_{\phi_\ell(x)}, \tag{4.15}$$

where

$$f_\ell = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_\ell} e_\ell(x_i).$$

The coefficients $\{f_\ell\}_{\ell=1}^{\infty}$ are square summable since

$$\|f\|_{\ell_2} = \left\| \sum_{i=1}^{m} \alpha_i \phi(x_i) \right\| \leq \sum_{i=1}^{m} |\alpha_i| \, \|\phi(x_i)\| < \infty.$$

The key point is that we need never explicitly compute the eigenfunctions e_ℓ or the eigenexpansion (4.10) to specify functions in the RKHS: we simply write our functions in terms of the kernels, as in (4.14).

4.2 Formal definitions

In this section, we cover the **reproducing property**, which is what makes a Hilbert space a reproducing kernel Hilbert space (RKHS). We next show that every reproducing kernel Hilbert space has a unique positive definite kernel, and vice-versa: this is the Moore-Aronszajn theorem.

Our discussion of the reproducing property follows [1, Ch. 1] and [11, Ch. 4]. We use the notation $f(\cdot)$ to indicate we consider the function itself, and not just the function evaluated at a particular point. For the kernel $k(x_i, \cdot)$, one argument is fixed at x_i, and the other is free (recall the kernel is symmetric).

Definition 15 (Reproducing kernel Hilbert space (first definition))**.** [1, p. 7] Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel*[8] of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if k satisfies

- $\forall x \in \mathcal{X}, \;\; k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \;\; \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

[8] We've deliberately used the same notation for the kernel as we did for positive definite kernels earlier. We will see in the next section that we are referring in both cases to the same object.

In particular, for any $x, y \in \mathcal{X}$,

$$k(x,y) = \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}. \tag{4.16}$$

Recall that a kernel is an inner product between feature maps: then $\phi(x) = k(\cdot,x)$ is a valid feature map (so every reproducing kernel is indeed a kernel in the sense of Definition (3)).

The reproducing property has an interesting consequence for functions in $\mathcal{H}$. We define δ_x to be the operator of evaluation at x, i.e.

$$\delta_x f = f(x) \quad \forall f \in \mathcal{H},\ x \in \mathcal{X}.$$

We then get the following *equivalent* definition for a reproducing kernel Hilbert space.

Definition 16 (Reproducing kernel Hilbert space (second definition))**.** [11, Definition 4.18] $\mathcal{H}$ is an RKHS if for all $x \in \mathcal{X}$, the evaluation operator δ_x is bounded: there exists a corresponding $\lambda_x \geq 0$ such that $\forall f \in \mathcal{H}$,

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}}$$

This definition means that when two functions are identical in the RHKS norm, they agree at every point:

$$|f(x) - g(x)| = |\delta_x (f-g)| \leq \lambda_x \|f-g\|_{\mathcal{H}} \quad \forall f, g \in \mathcal{H}.$$

This is a particularly useful property[9] if we're using RKHS functions to make predictions at a given point x, by optimizing over $f \in \mathcal{H}$. That these definitions are equivalent is shown in the following theorem.

Theorem 17 (Reproducing kernel equivalent to bounded δ_x)**.** *[1, Theorem 1] $\mathcal{H}$ is a reproducing kernel Hilbert space (i.e., its evaluation operators δ_x are bounded linear operators), if and only if $\mathcal{H}$ has a reproducing kernel.*

Proof. We only prove here that if $\mathcal{H}$ has a reproducing kernel, then δ_x is a bounded linear operator. The proof in the other direction is more complicated [11, Theorem 4.20], and will be covered in the advanced part of the course (briefly, it uses the Riesz representer theorem).

Given that a Hilbert space $\mathcal{H}$ has a reproducing kernel k with the reproducing property $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x)$, then

$$\begin{aligned} |\delta_x[f]| &= |f(x)| \\ &= |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}| \\ &\leq \|k(\cdot,x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \langle k(\cdot,x), k(\cdot,x)\rangle_{\mathcal{H}}^{1/2} \|f\|_{\mathcal{H}} \\ &= k(x,x)^{1/2} \|f\|_{\mathcal{H}} \end{aligned}$$

where the third line uses the Cauchy-Schwarz inequality. Consequently, $\delta_x : \mathcal{F} \to \mathbb{R}$ is a bounded linear operator where $\lambda_x = k(x,x)^{1/2}$. □

[9] This property certainly does not hold for all Hilbert spaces: for instance, it fails to hold on the set of square integrable functions $L_2(\mathcal{X})$.

Finally, the following theorem is very fundamental [1, Theorem 3 p. 19], [11, Theorem 4.21], and will be proved in the advanced part of the course:

Theorem 18 (Moore-Aronszajn)**.** *[1, Theorem 3] Every positive definite kernel k is associated with a unique RKHS $\mathcal{H}$.*

Note that the feature map is *not* unique (as we saw earlier): only the kernel is. Functions in the RKHS can be written as linear combinations of feature maps,

$$f(\cdot) := \sum_{i=1}^{m} \alpha_i k(x_i, \cdot),$$

as in Figure 4.5, as well as the limits of Cauchy sequences (where we can allow $m \to \infty$).

5 Application 1: Distance between means in feature space

Suppose we have two distributions p, q and we sample $(x_i)_{i=1}^m$ from p and $(y_i)_{i=1}^n$ from q. What is the distance between their means *in feature space*? This exercise illustrates that using the reproducing property, you can compute this distance without ever having to evaluate the feature map.

Answer:

$$\begin{aligned}
\left\| \frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(y_j) \right\|_{\mathcal{H}}^2 &= \left\langle \frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(y_j), \frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(y_j) \right\rangle_{\mathcal{H}} \\
&= \frac{1}{m^2}\left\langle \sum_{i=1}^{m}\phi(x_i), \sum_{i=1}^{m}\phi(x_i) \right\rangle + \dots \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, y_j).
\end{aligned}$$

What might this distance be useful for? In the case $\phi(x) = x$, we can use this statistic to distinguish distributions with different means. If we use the feature mapping $\phi(x) = [x\ x^2]$ we can distinguish both means and variances. More complex feature spaces permit us to distinguish increasingly complex features of the distributions. As we'll see in much more detail later in the course, there are kernels that can distinguish *any* two distributions [3, 10].

6 Application 2: Kernel PCA

This is one of the most famous kernel algorithms: see [7, 8].

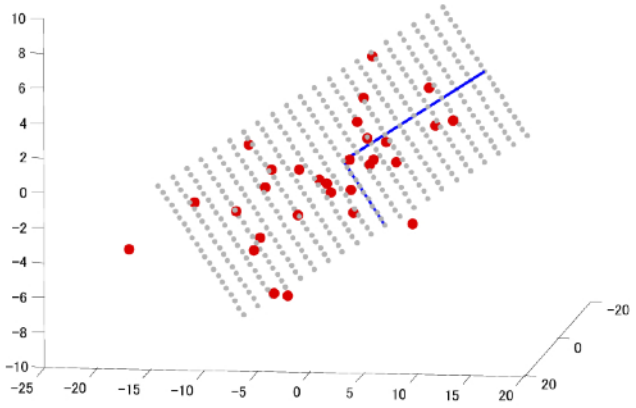

Figure 6.1: PCA in $\mathbb{R}^3$, for data in a two-dimensional subspace. The blue lines represent the first two principal directions, and the grey dots represent the 2-D plane in $\mathbb{R}^3$ on which the data lie (figure by Kenji Fukumizu).

6.1 Description of the algorithm

Goal of classical PCA: to find a d-dimensional subspace of a higher dimensional space (D-dimensional, $\mathbb{R}^D$) containing the directions of maximum variance. See Figure 6.1.

$$\begin{aligned} u_1 &= \arg\max_{\|u\|\leq 1} \frac{1}{n}\sum_{i=1}^{n}\left(u^\top\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)^2 \\ &= \arg\max_{\|u\|\leq 1} u^\top C u \end{aligned}$$

where

$$\begin{aligned} C &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^\top \\ &= \frac{1}{n} X H X^\top, \end{aligned}$$

where $X = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$, $H = I - n^{-1}\mathbf{1}_{n\times n}$, and $\mathbf{1}_{n\times n}$ is an $n \times n$ matrix of ones (note that $H = HH$, i.e. the matrix H is idempotent). We've looked at the first principal component, but all of the principal components u_i are the eigenvectors of the covariance matrix C (thus, each is orthogonal to all the previous ones). We have the eigenvalue equation

$$\lambda_i u_i = C u_i.$$

We now do this in feature space:

$$\begin{aligned} f_1 &= \arg\max_{\|f\|_{\mathcal{H}}\leq 1} \frac{1}{n}\sum_{i=1}^{n}\left(\left\langle f, \phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right\rangle_{\mathcal{H}}\right)^2 \\ &= \arg\max_{\|f\|_{\mathcal{H}}\leq 1} \mathrm{var}(f). \end{aligned}$$

First, observe that we can write

$$\begin{aligned} f_\ell &= \sum_{i=1}^{n}\alpha_{\ell i}\left(\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right), \\ &= \sum_{i=1}^{n}\alpha_{\ell i}\tilde{\phi}(x_i), \end{aligned}$$

since any component orthogonal to the span of $\tilde{\phi}(x_i) := \phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ vanishes when we take the inner product. The f are now elements of the feature space: if we were to use a Gaussian kernel, we could plot the *function* f by choosing the canonical feature map $\phi(x) = k(x, \cdot)$.

We can also define an infinite dimensional analog of the covariance:

$$\begin{aligned} C &= \frac{1}{n}\sum_{i=1}^{n}\left(\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)\otimes\left(\phi(x_i) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right), \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i) \end{aligned}$$

where we use the definition

$$(a\otimes b)c := \langle b, c\rangle_{\mathcal{H}}\, a \tag{6.1}$$

this is analogous to the outer product of finite dimensional vectors, $(ab^\top)c = a(b^\top c)$. Writing this, we get

$$f_\ell\lambda_\ell = Cf_\ell. \tag{6.2}$$

Let's look at the right hand side: to apply (6.1), we use

$$\begin{aligned} &\left\langle \tilde{\phi}(x_i), \sum_{j=1}^{n}\alpha_{\ell j}\tilde{\phi}(x_j)\right\rangle_{\mathcal{H}} \\ =&\sum_{j=1}^{n}\alpha_{\ell j}\tilde{k}(x_i, x_j), \end{aligned}$$

where $\tilde{k}(x_i, x_j)$ is the (i,j)th entry of the matrix $\tilde{K} := HKH$ (this is an easy exercise!). Thus,

$$Cf_\ell = \frac{1}{n}\sum_{i=1}^{n}\beta_{\ell i}\tilde{\phi}(x_i), \qquad \beta_{\ell i} = \sum_{j=1}^{n}\alpha_{\ell j}\tilde{k}(x_i, x_j).$$

We can now project both sides of (6.2) onto each of the centred mappings $\tilde{\phi}(x_q)$: this gives a set of equations which must all be satisifed to get an equivalent eigenproblem to (6.2). This gives

$$\left\langle \tilde{\phi}(x_q), \text{LHS} \right\rangle_{\mathcal{H}} = \lambda_\ell \left\langle \tilde{\phi}(x_q), f_\ell \right\rangle = \lambda_\ell \sum_{i=1}^{n} \alpha_{\ell i} \tilde{k}(x_q, x_i) \qquad \forall q \in \{1 \dots n\}$$

$$\left\langle \tilde{\phi}(x_q), \text{RHS} \right\rangle_{\mathcal{H}} = \left\langle \tilde{\phi}(x_q), C f_\ell \right\rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{k}(x_q, x_i) \left(\sum_{j=1}^{n} \alpha_{\ell j} \tilde{k}(x_i, x_j) \right) \qquad \forall q \in \{1 \dots n\}$$

Writing this as a matrix equation,

$$n\lambda_\ell \widetilde{K} \alpha_\ell = \widetilde{K}^2 \alpha_\ell,$$

or equivalently

$$n\lambda_\ell \alpha_\ell = \widetilde{K} \alpha_\ell. \tag{6.3}$$

Thus the α_ℓ are the eigenvectors of $\widetilde{K}$: it is not necessary to ever use the feature map $\phi(x_i)$ explicitly!

How do we ensure the eigenfunctions f have unit norm in feature space?

$$\begin{aligned}
&\|f\|_{\mathcal{H}}^2 \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i), \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_i \left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_i \tilde{k}(x_i, x_j) \\
&= \alpha^\top \widetilde{K} \alpha = n\lambda \alpha^\top \alpha = n\lambda \|\alpha\|^2.
\end{aligned}$$

Thus to re-normalise α such that $\|f\| = 1$, it suffices to replace

$$\alpha \leftarrow \alpha / \sqrt{n\lambda}$$

(assuming the original solutions to (6.3) have $\|\alpha\| = 1$).

How do you project a new point x^* onto the principal component f? Assuming f is properly normalised, the projection is

$$\begin{aligned}
P_f \phi(x^*) &= \langle \phi(x^*), f \rangle_{\mathcal{H}} f \\
&= \left(\sum_{j=1}^{n} \alpha_j \left\langle \phi(x^*), \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}} \right) \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i) \\
&= \left(\sum_{j=1}^{n} \alpha_j \left(k(x^*, x_j) - \frac{1}{n} \sum_{\ell=1}^{n} k(x^*, x_\ell) \right) \right) \sum_{i=1}^{n} \alpha_i \tilde{\phi}(x_i).
\end{aligned}$$

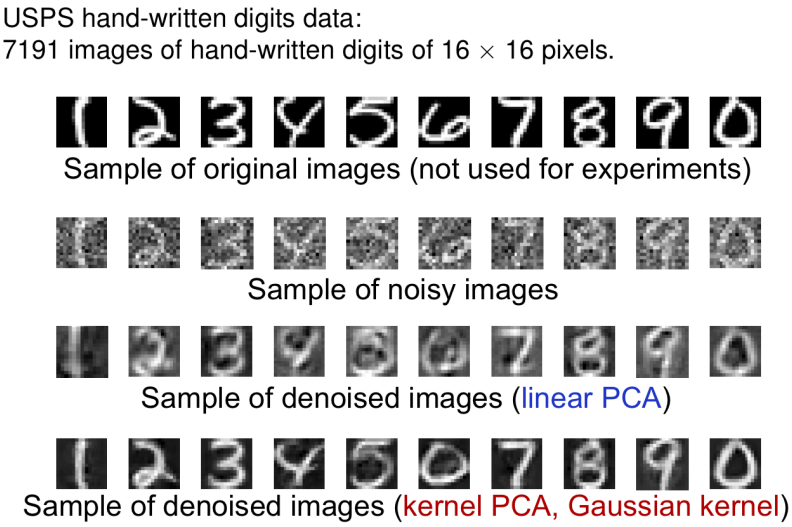

Figure 6.2: Hand-written digit denoising example (from Kenji Fukumizu's slides).

6.2 Example: image denoising

We consider the problem of denoising hand-written digits. Denote by

$$P_d\phi(x^*) = \quad P_{f_1}\phi(x^*) \quad + \ldots + \quad P_{f_d}\phi(x^*)$$

the projection of $\phi(x^*)$ onto one of the first d eigenvectors from kernel PCA (recall these are orthogonal). We define the nearest point $y \in \mathcal{X}$ to this feature space projection as the solution to the problem

$$y^* = \arg\min_{y \in \mathcal{X}} \|\phi(y) - P_d\phi(x^*)\|^2_{\mathcal{H}} .$$

In many cases, it will not be possible to reduce the squared error to zero, as there will be no single y^* corresponding to an exact solution. As in linear PCA, we can use the projection onto a subspace for denoising. By doing this in feature space, we can take into account the fact that data may not be distributed as a simple Gaussian, but can lie in a submanifold in input space, which nonlinear PCA can discover. See Figure 6.2.

7 Application 3: Ridge regression

In this section, we describe ridge regression. This is the algorithm used for the regression plots at the start of the document (Figure 2.3): it is very simple to implement and usually works quite well (except when the data have outliers, since it uses a squared loss).

7.1 A loss-based interpretation

7.1.1 Finite dimensional case

This discussion may be found in a number of sources. We draw from [9, Section 2.2]. We are given n training points in $\mathbb{R}^D$, which we arrange in a matrix $X = \left[\begin{array}{ccc} x_1 & \dots & x_n \end{array}\right] \in \mathbb{R}^{D\times n}$. To each of these points, there corresponds an output y_i, which we arrange in a column vector $y := \left[\begin{array}{ccc} y_1 & \dots & y_n \end{array}\right]^\top$. Define some $\lambda > 0$. Our goal is:

$$\begin{aligned} a^* &= \arg\min_{a\in\mathbb{R}^D} \left(\sum_{i=1}^n (y_i - x_i^\top a)^2 + \lambda \|a\|^2 \right) \\ &= \arg\min_{a\in\mathbb{R}^D} \left(\left\|y - X^\top a\right\|^2 + \lambda \|a\|^2 \right), \end{aligned}$$

where the second term $\lambda\|a\|^2$ is chosen to avoid problems in high dimensional spaces (see below). Expanding out the above term, we get

$$\begin{aligned} \left\|y - X^\top a\right\|^2 + \lambda\|a\|^2 &= y^\top y - 2y^\top X a + a^\top X X^\top a + \lambda a^\top a \\ &= y^\top y - 2y^\top X^\top a + a^\top \left(XX^\top + \lambda I\right) a = (*) \end{aligned}$$

Define $b = \left(XX^\top + \lambda I\right)^{1/2} a$, where the square root is well defined since the matrix is positive definite (it may be that $XX^\top$ is not invertible, for instance, when $D > n$, so adding λI ensures we can substitute $a = \left(XX^\top + \lambda I\right)^{-1/2} b$). Then

$$\begin{aligned} (*) &= y^\top y - 2y^\top X^\top \left(XX^\top + \lambda I\right)^{-1/2} b + b^\top b \\ &= y^\top y + \left\| \left(XX^\top + \lambda I\right)^{-1/2} Xy - b \right\|^2 - \left\| y^\top X^\top \left(XX^\top + \lambda I\right)^{-1/2} \right\|^2, \end{aligned}$$

where we complete the square. This is minimized when

$$\begin{aligned} b^* &= \left(XX^\top + \lambda I\right)^{-1/2} Xy \quad \text{or} \\ a^* &= \left(XX^\top + \lambda I\right)^{-1} Xy, \end{aligned}$$

which is the classic regularized least squares solution.[10]

[10] This proof differs from the usual derivation, which we give here for ease of reference (this is not the approach we use, since we are later going to extend our reasoning to feature spaces: derivatives in feature space also exist when the space is infinite dimensional, however for the purposes of ridge regression they can be avoided). We use [5, eqs. (61) and (73)]

$$\frac{\partial a^\top U a}{\partial a} = (U + U^\top)a, \qquad \frac{\partial v^\top a}{\partial a} = \frac{\partial a^\top v}{\partial a} = v,$$

Taking the derivative of the expanded expression $(*)$ and setting to zero,

$$\begin{aligned} \frac{\partial}{\partial a}\left(\left\|y - X^\top a\right\|^2 + \lambda\|a\|^2 \right) &= -2Xy + 2\left(XX^\top + \lambda I\right) a = 0, \\ a &= \left(XX^\top + \lambda I\right)^{-1} Xy. \end{aligned}$$

7.1.2 Finite dimensional case: more informative expression

We may rewrite this expression in a way that is more informative (and more easily kernelized). Assume without loss of generality that $D > n$ (this will be useful when we move to feature spaces, where D can be very large or even infinite). We can perform an SVD on X, i.e.

$$X = USV^\top,$$

where

$$U = \left[\begin{array}{ccc} u_1 & \dots & u_D \end{array} \right] \quad S = \left[\begin{array}{cc} \tilde{S} & 0 \\ 0 & 0 \end{array} \right] \quad V = \left[\begin{array}{cc} \tilde{V} & 0 \end{array} \right].$$

Here U is $D \times D$ and $U^\top U = UU^\top = I_D$ (the subscript denotes the size of the unit matrix), S is $D \times D$, where the top left diagonal $\tilde{S}$ has n non-zero entries, and V is $n \times D$, where only the first n columns are non-zero, and $\tilde{V}^\top \tilde{V} = \tilde{V}\tilde{V}^\top = I_n$.[11] Then

$$\begin{aligned} a^* &= \left(XX^\top + \lambda I_D\right)^{-1} Xy \\ &= \left(US^2U^\top + \lambda I_D\right)^{-1} USV^\top y \\ &= U\left(S^2 + \lambda I_D\right)^{-1} U^\top USV^\top y \\ &= U\left(S^2 + \lambda I_D\right)^{-1} SV^\top y \\ &= US\left(S^2 + \lambda I_D\right)^{-1} V^\top y \\ &= U\underbrace{SV^\top V}_{\text{(a)}}\left(S^2 + \lambda I_D\right)^{-1} V^\top y \\ &\underset{\text{(b)}}{=} X(X^\top X + \lambda I_n)^{-1} y \end{aligned} \tag{7.1}$$

Step (a) is allowed since both S and $V^\top V$ are non-zero in the same sized top-left block, and $V^\top V$ is just the unit matrix in that block. Step (b) occurs as follows

$$\begin{aligned} V\left(S^2 + \lambda I_D\right)^{-1} V^\top &= \left[\begin{array}{cc} \tilde{V} & 0 \end{array} \right] \left[\begin{array}{cc} \left(\tilde{S}^2 + \lambda I_n\right)^{-1} & 0 \\ 0 & (\lambda I_{D-n})^{-1} \end{array} \right] \left[\begin{array}{c} \tilde{V}^\top \\ 0 \end{array} \right] \\ &= \tilde{V}\left(\tilde{S}^2 + \lambda I_n\right)^{-1} \tilde{V}^\top \\ &= \left(X^\top X + \lambda I_n\right)^{-1}. \end{aligned}$$

What's interesting about this result is that $a^* = \sum_{i=1}^n \alpha_i^* x_i$, i.e. a is a weighted sum of columns of X. Again, one can obtain this result straightfor-

[11] Another more economical way to write the SVD would be

$$X = U\left[\begin{array}{c} \tilde{S} \\ 0 \end{array} \right] \tilde{V}^\top,$$

but as we'll see, we will need the larger form.

wardly by applying established linear algebra results: the proof here is informative, however, since we are explicitly demonstrating the steps we take, and hence we can be assured the same steps will still work even if D is infinite.[12]

7.1.3 Feature space case

We now consider the case where we use features $\phi(x_i)$ in the place of x_i:

$$a^* \quad = \quad \arg\min_{a\in\mathcal{H}} \left(\sum_{i=1}^{n} \left(y_i - \langle a, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|a\|_{\mathcal{H}}^2 \right).$$

We could consider a number of options: e.g. the polynomial features or sinusoidal features

$$\phi_p(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^\ell \end{bmatrix} \qquad \phi_s(x) = \begin{bmatrix} \sin x \\ \cos x \\ \sin 2x \\ \vdots \\ \cos \ell x \end{bmatrix}$$

In these cases, a is a vector of length ℓ giving weight to each of these features so as to find the mapping between x and y. We can also consider feature vectors of *infinite* length, as we discussed before.

It is straightforward to obtain the feature space solution of the ridge regression equation in the light of the previous section: with some cumbersome notation, write[13]

$$X = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix}.$$

All of the steps that led us to (7.1) then follow, where in particular

$$XX^\top = \sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i)$$

(using the notation (6.1) we introduced from kernel PCA), and

$$(X^\top X)_{ij} = \langle \phi(x_i), \phi(x_j)\rangle_{\mathcal{H}} = k(x_i, x_j).$$

[12] We could apply one of the many variants of the Woodbury identity [5, eqs. (147)]. If P and R are positive definite, then

$$\left(P^{-1} + B^\top R^{-1} B\right) B^\top R^{-1} = PB^\top \left(BPB^\top + R\right)^{-1}. \tag{7.2}$$

Setting $P = \lambda^{-1} I$, $B = X^\top$, and $R = I$, we get

$$\begin{aligned} a = \left(XX^\top + \lambda I\right)^{-1} Xy &= \lambda^{-1} X \left(\lambda^{-1} X^\top X + I\right) y \\ &= X \left(X^\top X + \lambda I\right) y. \end{aligned}$$

[13] For infinite dimensional feature spaces, the operator X still has a singular value decomposition - this will be covered later in the course.

Making these replacements, we get

$$\begin{aligned} a^* &= X(K+\lambda I_n)^{-1}y \\ &= \sum_{i=1}^{n} \alpha_i^* \phi(x_i) \qquad \alpha^* = (K+\lambda I_n)^{-1}y. \end{aligned}$$

Note that the proof becomes much easier if we *begin* with the knowledge that a is a linear combination of feature space mappings of points,[14]

$$a = \sum_{i=1}^{n} \alpha_i \phi(x_i).$$

Then

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \langle a, \phi(x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|a\|_{\mathcal{H}}^2 &= \|y - K\alpha\|^2 + \lambda \alpha^\top K \alpha \\ &= y^\top y - 2y^\top K\alpha + \alpha^\top \left(K^2 + \lambda K\right)\alpha \end{aligned}$$

Differentiating wrt α and setting this to zero, we get

$$\alpha^* = (K+\lambda I_n)^{-1}y$$

as before.

7.2 Link of RKHS norm with smoothness of regression function

What does $\|f\|_{\mathcal{H}}^2$ have to do with smoothing? We illustrate this with two examples, taken from earlier in the notes. Recall from Section 4.1.3 that functions in the Gaussian RKHS take the form

$$f(x) = \sum_{\ell=1}^{\infty} \underbrace{f_\ell \sqrt{\lambda_\ell}}_{\hat{f}_\ell} e_\ell(x), \qquad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell},$$

where the eigenfunctions $e_\ell(x)$ were illustrated in Figure (4.4), and satisfy the orthonormality condition (4.8) for the measure (4.9). The constraint $\|f\|_{\mathcal{H}}^2 < \infty$ means that the $\hat{f}_\ell^2$ must decay faster than λ_ℓ with increasing ℓ. In other words, basis functions $e_\ell(x)$ with larger ℓ are given less weight: these are the non-smooth functions.

The same effect can be seen if we use the feature space in Section 4.1.2. Recall that functions on the periodic domain $[-\pi, \pi]$ have the representation

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x).$$

[14] This is a specific instance of the representer theorem, which we will encounter later.

Again,

$$\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{l=-\infty}^{\infty} \frac{\left|\hat{f}_\ell\right|^2}{\hat{k}_\ell}.$$

This means $\left|\hat{f}_\ell\right|^2$ must decay faster than $\hat{k}_\ell$ for the norm to be finite.[15] This serves to suppress the terms $\exp(\imath\ell x)$ for large ℓ, which are the non-smooth terms.

7.3 Model selection

In kernel ridge regression, we have control over two things: the kernel we use, and the weight λ. The kernel used controls the smoothness of the class of functions we consider. The weight λ controls the tradeoff between function smoothness and fitting error. We now look at these properties more closely, doing kernel ridge regression with a Gaussian kernel,

$$k(x, y) = \exp\left(\frac{-\|x - y\|^2}{\sigma}\right).$$

From Figure 7.1, we see that too large a λ prioritises smoothness over getting a small prediction error on the points, resulting in a very smooth function which barely follows the shape of the underlying data - in other words, we are **underfitting**. Too small a λ gives too much priority to fitting small fluctuations in the data due to noise, at the expense of smoothness: in this case, we are **overfitting**. Finally, an apparently good choice is $\lambda = 0.1$, where the regression curve fits the underlying trend without being overly influenced by noise.

Figure 7.2 shows how the kernel width σ affects the fit of ridge regression. Too large a σ results in underfitting: the regression function is too smooth. Too small a σ results in overfitting. There is some overlap in the effect on prediction quality of σ and λ.

How do we choose λ and σ, and how do we evaluate the resulting performance of our learning algorithm? One commonly used approach is to combine m-fold cross-validation and a held-out test set. See Algorithm 1.

7.4 A Bayesian interpretation

The Bayesian interpretation of ridge regression can be found in [6, Chapter 2]. Advantage: also get uncertainty estimate in the prediction.

8 Acknowledgements

Thanks to Gergo Bohner, Peter Dayan, Agnieszka Grabska-Barwinska, Wittawat Jitkrittum, Peter Latham, Arian Maleki, Kirsty McNaughton, Sam Pat-

[15] The rate of decay of $\hat{k}_l$ will depend on the properties of the kernel. Some relevant results may be found at `http://en.wikipedia.org/wiki/Convergence_of_Fourier_series`

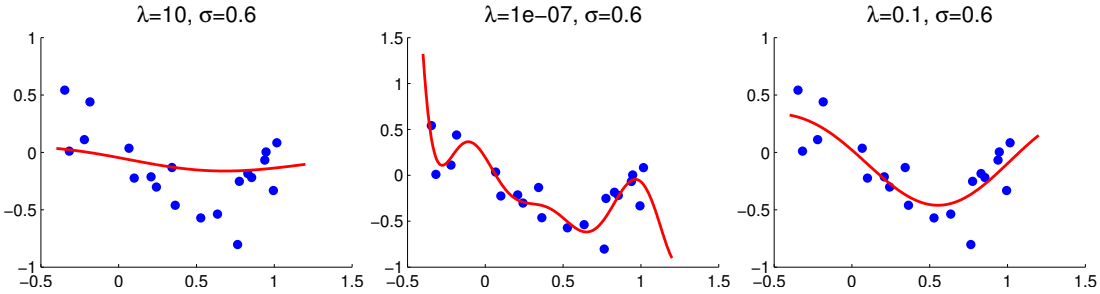

Figure 7.1: Effect of choice of λ on the fit of ridge regression.

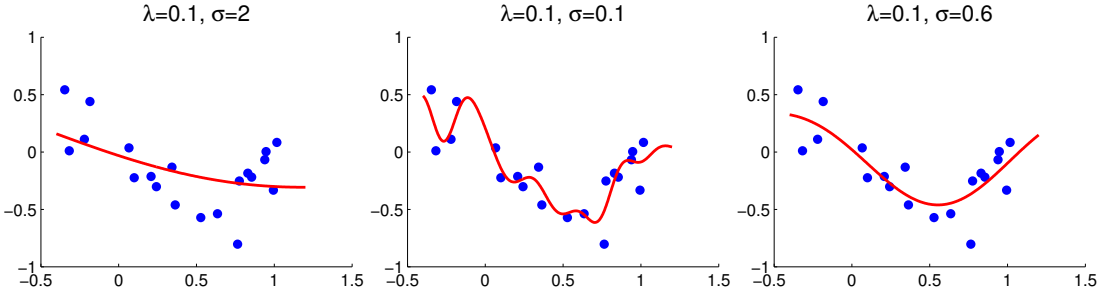

Figure 7.2: Effect of choice of σ on the fit of ridge regression.

Algorithm 1 m-fold cross validation and held-out test set.

1. Start with a dataset $Z := X, Y$, where X is a matrix with n columns, corresponding to the n training points, and Y is a vector having n rows. We split this into a training set of size n_{tr} and a test set of size $n_{\text{te}} = 1 - n_{\text{tr}}$.
2. Break the trainining set into m equally sized chunks, each of size $n_{\text{val}} = n_{\text{tr}}/m$. Call these $X_{\text{val},i}, Y_{\text{val},i}$ for $i \in \{1, \ldots, m\}$
3. For each λ, σ pair
   (a) For each $X_{\text{val},i}, Y_{\text{val},i}$
      i. Train the ridge regression on the remaining trainining set data $X_{\text{tr}} \setminus X_{\text{val},i}$ and $Y_{\text{tr}} \setminus Y_{\text{val},i}$,
      ii. Evaluate its error on the validation data $X_{\text{val},i}, Y_{\text{val},i}$
   (b) Average the errors on the validation sets to get the average validation error for λ, σ.
4. Choose λ^*, σ^* with the lowest average validation error
5. Measure the performance on the test set $X_{\text{te}}, Y_{\text{te}}$.

Table 1: Fourier series relations in 1-D.

Description of rule	Input space	Frequency space
Shift	$f(x - x_0)$	$\tilde{f}_l \exp(-\imath l (2\pi/T) x_0)$
Input real	$f^*(x) = f(x)$	$\tilde{f}_l = -\tilde{f}_l^*$
Input even, real	$f^*(x) = f(x)$, $f(-x) = f(x)$	$\tilde{f}_l = -\tilde{f}_l^*$
Scaling	$f(ax)$	T changes accordingly
Differentiation	$\frac{d}{dx} f(x)$	$\imath l (2\pi/T) \tilde{f}_l$
Parseval's theorem	$\int_{-T/2}^{T/2} f(x) g^*(x) dx$	$\sum_{k=-\infty}^{\infty} \tilde{f}_l \tilde{g}_l^*$

terson, Dino Sejdinovic, and Yingjian Wang for providing feedback on the notes, and correcting errors.

A The Fourier series on $\left[-\frac{T}{2}, \frac{T}{2}\right]$ with periodic boundary conditions

We consider the case in which $f(x)$ is periodic with period T, so that we need only specify $f(x) : \left[-\frac{T}{2}, \frac{T}{2}\right] \to \mathbb{R}$. In this case, we obtain the Fourier series expansion

$$\tilde{f}_l = \frac{1}{T} \int_{-T/2}^{T/2} f(x) \exp\left(-\imath l x \frac{2\pi}{T}\right) dx = \frac{1}{T} \tilde{f}\left(l \frac{2\pi}{T}\right), \tag{A.1}$$

such that

$$f(x) = \sum_{l=-\infty}^{\infty} \tilde{f}_l \exp\left(\imath l x \frac{2\pi}{T}\right). \tag{A.2}$$

Thus the $\tilde{f}_l$ represent the Fourier transform at frequencies $\omega = l\frac{2\pi}{T}$, scaled by T^{-1}. We document a number of useful Fourier series relations in Table 1.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[2] Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: On the bias–variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, October 2002.

[3] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems 15*, pages 513–520, Cambridge, MA, 2007. MIT Press.

[4] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, February 2002.

[5] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. Version 20081110.

[6] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[7] B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.

[8] Bernhard Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[9] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

[10] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[11] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, 2008.

[12] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

Notes on mean embeddings and covariance operators

Arthur Gretton

January 11, 2020

1 Introduction

This note contains more detailed proofs of certain results in the lecture notes on mean embeddings and covariance operators. The notes are not as complete as for lectures 1 and 2, but cover only the trickier concepts. Please let me know if there are any further parts you'd like clarified, and I'll add them to the note.

2 Mean embeddings

2.1 Proof that the mean embedding exists via Riesz

For finite dimensional feature spaces, we can define expectations in terms of inner products.

$$\phi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

Then

$$f(x) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} = \langle f, \phi(x) \rangle_{\mathcal{F}} .$$

Consider random variable $\mathsf{x} \sim \mathbf{P}$

$$\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) = \mathbf{E}_{\mathbf{P}} \left(\begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathsf{x} \\ \mathsf{x}^2 \end{bmatrix} \right) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathbf{E}_{\mathbf{P}} \mathsf{x} \\ \mathbf{E}_{\mathbf{P}} (\mathsf{x}^2) \end{bmatrix} =: \begin{bmatrix} a \\ b \end{bmatrix}^\top \mu_{\mathbf{P}}.$$

Does this reasoning translate to infinite dimensions?

Definition 1 (Bounded operator)**.** A linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$Af \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

We prove via Riesz that the mean embedding exists, and that it takes the form of the expectation of the canonical map.

Theorem 2. *[Riesz representation]In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,*

$$Af = \langle f, g_A \rangle_{\mathcal{F}}$$

Now we establish the existence of the mean embedding.

Lemma 3 (Existence of mean embedding)**.** *If $\mathbf{E}_{\mathsf{P}} \sqrt{k(\mathsf{x}, \mathsf{x})} < \infty$ then $\mu_{\mathsf{P}} \in \mathcal{F}$.*

Proof. The linear operator $T_{\mathsf{P}} f := \mathbf{E}_{\mathsf{P}} f(\mathsf{x})$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_{\mathsf{P}} f| = |\mathbf{E}_{\mathsf{P}} f(\mathsf{x})| \underset{(a)}{\leq} \mathbf{E}_{\mathsf{P}} |f(\mathsf{x})| = \mathbf{E}_{\mathsf{P}} |\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}}| \leq \mathbf{E}_{\mathsf{P}} \left(\sqrt{k(\mathsf{x}, \mathsf{x})} \|f\|_{\mathcal{F}} \right),$$

where in (a) we use Jensen's inequality. Hence by the Riesz representer theorem [6, Theorem II.4], there exists a $\mu_{\mathsf{P}} \in \mathcal{F}$ such that $T_{\mathsf{P}} f = \langle f, \mu_{\mathsf{P}} \rangle_{\mathcal{F}}$. □

If we set $f = \phi(x) = k(x, \cdot)$, we obtain $\mu_{\mathsf{P}}(x) = \langle \mu_p, k(x, \cdot) \rangle = \mathbf{E}_{\mathsf{P}} k(x, \mathsf{x})$: in other words, the mean embedding of the distribution P is the expecation under P of the canonical feature map.

2.2 Proof that MMD injective for universal kernel

First, it is clear that $\mathsf{P} = \mathsf{Q}$ implies MMD $\{\mathsf{P}, \mathsf{Q}; F\}$ is zero. We now prove the converse. By the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ there exists a $g \in \mathcal{F}$ such that

$$\|f - g\|_{\infty} \leq \epsilon.$$

We will need [2, Lemma 9.3.2]:

Lemma 4. *Let $(\mathcal{X}, d)$ be a metric space, and let P, Q be two Borel probability measures defined on $\mathcal{X}$, where we define the random variables $\mathsf{x} \sim \mathsf{P}$ and $\mathsf{y} \sim \mathsf{Q}$. Then $\mathsf{P} = \mathsf{Q}$ if and only if $\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) = \mathbf{E}_{\mathsf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.*

We now use these two results to formulate a proof. We begin with the expansion

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq |\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| + |\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y})| + |\mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})|.$$

The first and third terms satisfy

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| \leq \mathbf{E}_{\mathsf{P}} |f(\mathsf{x}) - g(\mathsf{x})| \leq \epsilon.$$

Next, write

$$\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) = \langle g, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle = 0,$$

since MMD $\{\mathsf{P}, \mathsf{Q}; F\} = 0$ implies $\mu_{\mathsf{P}} = \mu_{\mathsf{Q}}$. Hence

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq 2\epsilon$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies $\mathsf{P} = \mathsf{Q}$ by Lemma 4.

3 Covariance operators

One of the most important and widely used tools in RKHS theory is the covariance operator: this is an infinite dimensional analog to the covariance matrix. This forms the backbone of kernel PCA, the kernel Fisher discriminant, kernel partial least squares, the kernel canonical correlation, and so on.

In this note, we describe the Hilbert space of Hilbert-Schmidt operators. We then introduce the covariance operator, demonstrate it is Hilbert-Schmidt, and express it in terms of kernel functions.

3.1 Hilbert-Schmidt operators

This discussion is based on [9, Section 2.1] and [8, Section A.5.2].

Let $\mathcal{F}$ and $\mathcal{G}$ be separable Hilbert spaces. Define $(e_i)_{i\in I}$ to be an orthonormal basis for $\mathcal{F}$, and $(f_j)_{j\in J}$ to be an orthonormal basis for $\mathcal{G}$. The index sets I, J are assumed to be either finite our countably infinite.[1] Define two compact linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$. Define the Hilbert-Schmidt norm of the operators L, M to be

$$\begin{aligned} \|L\|_{\mathrm{HS}}^2 &= \sum_{j\in J} \|Lf_j\|_{\mathcal{F}}^2 \\ &= \sum_{i\in I}\sum_{j\in J} \left|\langle Lf_j, e_i\rangle_{\mathcal{F}}\right|^2, \end{aligned} \tag{3.1}$$

where we use Parseval's indentity on each of the norms in the first sum. The operator L is Hilbert-Schmidt when this norm is finite.

The Hilbert-Schmidt operators mapping from $\mathcal{G}$ to $\mathcal{F}$ form a Hilbert space, written $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{j\in J} \langle Lf_j, Mf_j\rangle_{\mathcal{F}}, \tag{3.2}$$

which is independent of the orthonormal basis chosen. It is clear the norm (3.1) is recovered from this inner product. Another form for this inner product is

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{i\in I}\sum_{j\in J} \langle Lf_j, e_i\rangle_{\mathcal{F}} \langle Mf_j, e_i\rangle_{\mathcal{F}}. \tag{3.3}$$

Proof. Since any element of $\mathcal{F}$ can be expanded in terms of its orthonormal basis, we have that this holds in the specific case of the mapping of f_j by L or M,

$$Lf_j = \sum_{i\in I} \alpha_i^{(j)} e_i \qquad Mf_j = \sum_{i'\in I} \beta_{i'}^{(j)} e_{i'}. \tag{3.4}$$

[1] Recall that a Hilbert space has a countable orthonormal basis if and only if it is separable: that is, it has a countable dense subset [6, p. 47].

Substituting these into (3.2), we obtain

$$\begin{aligned}\langle L, M\rangle_{\mathrm{HS}} &= \sum_{j\in J}\left\langle \sum_{i\in I}\alpha_i^{(j)} e_i, \sum_{i'\in I}\beta_{i'}^{(j)} e_{i'}\right\rangle_{\mathcal{F}} \\ &= \sum_{i\in I}\sum_{j\in J}\alpha_i^{(j)}\beta_i^{(j)}.\end{aligned}$$

We obtain the identical result when we substitute (3.4) into (3.3). □

3.2 Rank-one operators, tensor product space

Given $b \in \mathcal{G}$ and $a \in \mathcal{F}$, we define the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$,

$$(b \otimes a) f \;\mapsto\; \langle f, a\rangle_{\mathcal{F}}\, b. \tag{3.5}$$

This is a generalization of the standard outer product in linear algebra, $(ba^\top)f = (a^\top f)b$, if all three of a, b, f were vectors. First, is this operator Hilbert-Schmidt? We compute its norm according to (3.1),

$$\begin{aligned}\|a \otimes b\|_{\mathrm{HS}}^2 &= \sum_{j\in J}\|(a\otimes b) f_j\|_{\mathcal{F}}^2 \\ &= \sum_{j\in J}\left\|a\,\langle b, f_j\rangle_{\mathcal{G}}\right\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j\in J}\left|\langle b, f_j\rangle_{\mathcal{G}}\right|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2,\end{aligned} \tag{3.6}$$

where we use Parseval's identity. Thus, the operator is Hilbert-Schmidt.

Given a second Hilbert-Schmidt operator $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, we have the result:

$$\langle L, a\otimes b\rangle_{\mathrm{HS}} = \langle a, Lb\rangle_{\mathcal{F}} \tag{3.7}$$

A particular instance of this result is

$$\langle u\otimes v, a\otimes b\rangle_{\mathrm{HS}} = \langle u, a\rangle_{\mathcal{F}}\,\langle b, v\rangle_{\mathcal{G}}\,. \tag{3.8}$$

Proof. The key result we use is the expansion of b in terms of the orthonormal basis, $b = \sum_{j\in J}\langle b, f_j\rangle_{\mathcal{G}}\, f_j$. Then

$$\begin{aligned}\langle a, Lb\rangle &= \left\langle a, L\left(\sum_j \langle b, f_j\rangle_{\mathcal{G}} f_j\right)\right\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}}\,\langle a, Lf_j\rangle_{\mathcal{F}}\end{aligned}$$

and

$$\begin{aligned}\langle a \otimes b, L\rangle_{\mathrm{HS}} &= \sum_j \langle Lf_j, (a \otimes b) f_j\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}} \langle Lf_j, a\rangle_{\mathcal{F}}.\end{aligned}$$

To show (3.8), we simply substitute $u \otimes v$ for L above, and then apply the definition (3.5),

$$\begin{aligned}\langle u \otimes v, a \otimes b\rangle_{\mathrm{HS}} \langle a, (u \otimes v)\, b\rangle_{\mathcal{F}} \\ = \langle u, a\rangle_{\mathcal{F}} \langle b, v\rangle_{\mathcal{G}}\end{aligned}$$

□

3.3 Cross-covariance operator

In this section, we define the cross-covariance operator, in the case where $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert spaces with respective kernels k and l, and feature maps ϕ and ψ. This is a generalization of the cross-covariance matrix to infinite dimensional feature spaces. The results we want are feature space analogues to:

$$\widetilde{C}_{XY} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right) \qquad f^\top \widetilde{C}_{XY} g = \mathbf{E}_{\mathsf{x}\mathsf{y}}\left[\left(f^\top \mathsf{x}\right)\left(g^\top \mathsf{y}\right)\right],$$

where we use the notation $\widetilde{C}_{XY}$ to denote a covariance operator without centering. The corresponding centered covariance is

$$C_{XY} := \widetilde{C}_{XY} - \mu_X \mu_Y^\top,$$

where $\mu_X := \mathbf{E}(\mathsf{x})$ and $\mu_Y := \mathbf{E}(\mathsf{y})$We now describe how we can get these results in feature space.

The cross product $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ is a random variable in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$: use the result in [9, p. 265] that for all $A \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, the linear form $\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}}$ is measurable. For the expectation of this random variable to exist (and to be an element of $\mathrm{HS}(\mathcal{G}, \mathcal{F})$), we require the expected norm of $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ to be bounded: in other words, $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$. Given the expectation exists, and writing it $\widetilde{C}_{XY}$, then this expectation is the unique element satisfying

$$\left\langle \widetilde{C}_{XY}, A\right\rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \tag{3.9}$$

Proof. The operator

$$\begin{aligned} T_{\mathsf{x}\mathsf{y}} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \mathbb{R} \\ A &\mapsto \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \end{aligned}$$

is bounded when $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$, since by applying first Jensen's inequality, then Cauchy-Schwarz,

$$
\begin{aligned}
|\mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| &\leq \mathbf{E}_{\mathsf{x},\mathsf{y}} |\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| \\
&\leq \|A\|_{\mathrm{HS}} \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right).
\end{aligned}
$$

Thus by the Riesz representer theorem (Theorem (2)), the covariance operator (3.9) exists. We can make a further simplification to the condition: substituting (3.6), we get the requirement

$$
\begin{aligned}
\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x})\|_{\mathcal{F}} \|\phi(\mathsf{y})\|_{\mathcal{G}} \right) \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\sqrt{k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y})} \right) < \infty.
\end{aligned}
$$

We could also use the weaker condition $\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y}) \right)$, which is implied from the above by Jensen's inequality. □

We now use the particular element $f \otimes g$. Combining (3.7) and (3.9), we have the result

$$
\begin{aligned}
\left\langle f, \widetilde{C}_{XY} g \right\rangle_{\mathcal{F}} &= \left\langle \widetilde{C}_{XY}, f \otimes g \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), f \otimes g \rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}} \langle g, \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[f(\mathsf{x}) g(\mathsf{y}) \right] = \mathrm{cov}(f, g).
\end{aligned}
$$

What does this operator look like? To see this, we apply it to $k(x, \cdot) l(y, \cdot)$ (just as we plotted the mean embedding by evaluating it on $k(x, \cdot)$).

We are given an i.i.d. sample from $\mathbf{P} = \mathbf{P}_{\mathsf{x}} \mathbf{P}_{\mathsf{y}}$, written $\mathbf{z} := ((x_1, y_1) \dots (x_n, y_n))$. Write the empirical

$$
\widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y,
$$

where we have now included the *centering terms* $\hat{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$. With some algebra, this can be written

$$
\widehat{C}_{XY} = \frac{1}{n} X H Y^\top,
$$

where $H = I_n - n^{-1} \mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix of ones, and

$$
X = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} \psi(y_1) & \dots & \psi(y_n) \end{bmatrix}.
$$

Define the kernel matrices

$$
K_{ij} = \left(X^\top X \right)_{ij} = k(x_i, x_j) \qquad L_{ij} = l(y_i, y_j),
$$

and the kernel matrices between centred variables,

$$
\widetilde{K} = HKH \qquad \widetilde{L} = HLH
$$

(exercise: prove that the above are kernel matrices for the variables centred in feature space).

4 Using the covariance operator to detect dependence

There are two measures of dependence we consider: the constrained covariance (COCO), which is the largest singular value of the covariance operator, and the Hilbert-Schmidt Independence Criterion, which is its Hilbert-Schmidt norm.

4.1 Empirical COCO and proof

We now derive the functions satisying

$$\begin{aligned} \text{maximize} \quad & \left\langle g, \widehat{C}_{XY} f \right\rangle_{\mathcal{G}} \\ \text{subject to} \quad & \|f\|_{\mathcal{F}} = 1 \qquad (4.1) \\ & \|g\|_{\mathcal{G}} = 1 \qquad (4.2) \end{aligned}$$

We assume that

$$f = \sum_{i=1}^{n} \alpha_i \left[\phi(x_i) - \hat{\mu}_x\right] = XH\alpha \qquad g = \sum_{j=1}^{n} \beta_i \left[\psi(y_i) - \hat{\mu}_y\right] = YH\beta,$$

where

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \qquad \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i).$$

The associated Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = -f^\top \widehat{C}_{XY} g + \frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right),$$

where we have negated the covariance to make it a minimization problem (for consistency with the optimisation lecture later in the course), and we divide the Lagrange multipliers by 2 to simplify the discussion later . We now write this in terms of α and β:

$$\begin{aligned} f^\top \widehat{C}_{XY} g &= \frac{1}{n}\alpha^\top H X^\top \left(XHY^\top\right) YH\beta \\ &= \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta, \end{aligned}$$

where we note that $H = HH$. Similarly

$$\|f\|_{\mathcal{F}}^2 = \alpha^\top HXX^\top H\alpha = \alpha^\top \widetilde{K}\alpha.$$

Substituting these into the Lagrangian, we get a new optimization in terms of α and β,

$$\mathcal{L}(\alpha, \beta, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right). \qquad (4.3)$$

We must minimise this wrt the primal variables α, β. Differentiating wrt α and β and setting the resulting expressions to zero,[2] we obtain

$$-\frac{1}{n}\widetilde{K}\widetilde{L}\beta + \lambda\widetilde{K}\alpha = 0 \tag{4.4}$$

$$-\frac{1}{n}\widetilde{L}\widetilde{K}\alpha + \gamma\widetilde{L}\beta = 0 \tag{4.5}$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta = \lambda\alpha^\top\widetilde{K}\alpha$$
$$\frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta$$

Subtracting the first expression from the second, we get

$$\lambda\alpha^\top\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta.$$

Thus for $\lambda \neq 0$ and $\gamma \neq 0$, we conclude that $\lambda = \gamma$. Making this replacement in (4.4) and (4.5), we must find the largest[3] γ that solves the following expression wrt α, β:

$$\begin{bmatrix} 0 & \frac{1}{n}\widetilde{K}\widetilde{L} \\ \frac{1}{n}\widetilde{L}\widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{4.6}$$

This is a generalized eigenvalue problem, and can be solved straightforwardly in Matlab. The maximum eigenvalue is indeed COCO: at the solution, $\alpha^\top\widetilde{K}\alpha = 1$ and $\beta^\top\widetilde{L}\beta = 1$, hence the two norm terms in the Lagrangian (4.3) vanish.[4]

[2] We use [5, eqs. (61) and (73)]

$$\frac{\partial a^\top U a}{\partial a} = (U + U^\top)a, \qquad \frac{\partial v^\top a}{\partial a} = \frac{\partial a^\top v}{\partial a} = v.$$

[3] Given $\lambda = \gamma$, the system of equations (4.4) and (4.5) becomes:

$$\frac{1}{n}\widetilde{K}\widetilde{L}\beta = \gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = \gamma\widetilde{L}\beta$$

However we also get a valid solution by switching $\check{\beta} := -\beta$,

$$\frac{1}{n}\widetilde{K}\widetilde{L}\check{\beta} = -\gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = -\gamma\widetilde{L}\check{\beta}$$

In other words, the solutions γ of the generalised eigenvalue problem (4.6) come in pairs $\pm\gamma$, depending on the relative sign of α and β.

[4] For a more roundabout way of reaching the same conclusion: pre-multiply (4.6) by $[\alpha^\top \beta^\top]$ to get the system of equations

$$\begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta \\ \frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha \end{bmatrix} = \gamma \begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\alpha \\ \frac{1}{n}\beta^\top\widetilde{L}\beta \end{bmatrix} = \gamma \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where in the final line we substitute the constraints from (4.1).

4.2 The Hilbert-Schmidt Independence Criterion

4.2.1 Population expression

What is the Hilbert-Schmidt norm of the covariance operator?[5] Consider the centered, squared norm of the RKHS covariance operator,

$$\begin{aligned} HSIC^2(\mathcal{F}, \mathcal{G}, P_{xy}) &= \|\widetilde{C}_{XY} - \mu_X \otimes \mu_Y\|^2_{\text{HS}} \\ &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} - 2 \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}}, \end{aligned}$$

where $\widetilde{C}_{XY}$ is the uncentered covariance operator defined in (3.9). There are three terms in the expansion.

To obtain the first term, we apply (3.9) twice, denoting by $(\mathsf{x}', \mathsf{y}')$ an independent copy of the pair of variables (x, y),

$$\begin{aligned} \|\widetilde{C}_{XY}\|^2_{\text{HS}} &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \phi(\mathsf{x}') \otimes \psi(\mathsf{y}') \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}), [\phi(\mathsf{x}') \otimes \psi(\mathsf{y}')] \psi(\mathsf{y}) \rangle_{\mathcal{F}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left[\langle \phi(\mathsf{x}), \phi(\mathsf{x}') \rangle_{\mathcal{F}} \langle \psi(\mathsf{y}'), \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}') \\ &=: A \end{aligned}$$

Similar reasoning can be used to show

$$\begin{aligned} \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} &= \langle \mu_X, \mu_X \rangle_{\mathcal{F}} \langle \mu_Y, \mu_Y \rangle_{\mathcal{G}} \\ &= \mathbf{E}_{\mathsf{x}\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \\ &=: D, \end{aligned}$$

and for the cross-terms,

$$\begin{aligned} \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}} &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \mu_X \otimes \mu_Y \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\langle \phi(\mathsf{x}), \mu_X \rangle_{\mathcal{F}} \langle \phi(\mathsf{y}), \mu_Y \rangle_{\mathcal{G}} \right) \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\mathbf{E}_{\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \right) \\ &=: B. \end{aligned}$$

[5] Other norms of the operator may also be used in determining dependence, e.g. the spectral norm from the previous section. Another statistic on the kernel spectrum is the Kernel Mutual Information, which is an upper bound on the true mutual information near independence, but is otherwise difficult to interpret [4]. One can also define independence statistics on the correlation operator [1], which may be better behaved for small sample sizes, although the asymptotic behavior is harder to analyze.

4.2.2 Biased estimate

A biased estimate of HSIC was given in [3]. We observe a sample $Z := \{(x_1, y_1), \dots (x_n, y_n)\}$ drawn independently and identically from P_{xy}, we wish to obtain empirical expressions for HSIC,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \widehat{A} - 2\widehat{B} + \widehat{D}.$$

A direct approach would be to replace the population uncentred covariance operator $\widetilde{C}_{XY}$ with an empirical counterpart,

$$\check{C}_{XY} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i),$$

and the population mean embeddings with their respective empirical estimates,

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \qquad \hat{\mu}_y = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i),$$

however the resulting estimates are biased (we will show the amount of bias in the next section). The first term is

$$\begin{aligned} \widehat{A}_b = \left\| \check{C}_{XY} \right\|^2 &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} l_{ij} = \frac{1}{n^2} \text{tr}(KL), \end{aligned}$$

we use the shorthand $k_{ij} = k(x_i, x_j)$, and the subscript b to denote a biased estimate. The expression is not computationally efficient, and is written this way for later use - in practice, we would never take the matrix product if the intent was then to compute the trace. Next,

$$\begin{aligned} \widehat{B}_b = \left\langle \check{C}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y \right\rangle &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \otimes \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(x_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \right\rangle_{F} \left\langle \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\mathcal{G}} \\ &= \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{q=1}^{n} k_{ij} l_{iq} \\ &= \frac{1}{n^3} 1_n^\top KL 1_n = \frac{1}{n^3} 1_n^\top LK 1_n \end{aligned}$$

(we will use both forms to get our final biased estimate of HSIC), and

$$\begin{aligned}\widehat{D}_b = \langle \hat{\mu}_X \otimes \hat{\mu}_Y, \hat{\mu}_X \otimes \hat{\mu}_Y \rangle &= \left\langle \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right), \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^4}\left(\sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j)\right)\left(\sum_{i=1}^n \sum_{j=1}^n l(y_i, y_j)\right) \\ &= \frac{1}{n^4}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\end{aligned}$$

We now combine these terms, to obtain the biased estimate

$$\begin{aligned}HSIC_b^2(\mathcal{F}, \mathcal{G}, Z) &= \frac{1}{n^2}\left(\text{tr}(KL) - \frac{2}{n} 1_n^\top KL 1_n + \frac{1}{n^2}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\right) \\ &= \frac{1}{n^2}\left[\text{tr}(KL) - \frac{1}{n}\text{tr}\left(1_n 1_n^\top KL\right) - \frac{1}{n}\text{tr}\left(K 1_n 1_n^\top L\right) + \frac{1}{n^2}\text{tr}\left(1_n 1_n^\top K 1_n 1_n^\top L\right)\right] \\ &= \frac{1}{n^2}\text{tr}\left[\left(I - \frac{1}{n} 1_n 1_n^\top\right) K \left(I - \frac{1}{n} 1_n 1_n^\top\right) L\right] \\ &= \frac{1}{n^2}\text{tr}(KHLH)\end{aligned}$$

where we define

$$H := I - \frac{1}{n} 1_n 1_n^\top$$

as a centering matrix (when pre-multiplied by a matrix it centers the rows; when post-multiplied, it centers the columns).

4.2.3 Unbiased estimate

An unbiased estimate of $A := \|\widetilde{C}_{XY}\|_{\text{HS}}^2$ is

$$\widehat{A} := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k_{ij} l_{ij} = \frac{1}{(n)_2} \sum_{(i,j) \in \mathbf{i}_2^n} k_{ij} l_{ij},$$

where $\mathbf{i}_p^n$ is the set of all p-tuples drawn from $\{1, \ldots, n\}$, and

$$(n)_p = \frac{n!}{(n-p)!} = \frac{1}{n(n-1)\ldots(n-p+1)}.$$

Note that $\mathbf{E}(\widehat{A}) = \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}')$, which is not true of the biased expression (which does not properly treat the independent copies x' of x and y' of

y). The difference between the biased and unbiased estimates is

$$\begin{aligned}\widehat{A}_b - \widehat{A} &= \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}l_{ij} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} k_{ii}l_{ii} + \left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right)\left(\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right) \\ &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} k_{ii}l_{ii} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right),\end{aligned}$$

thus the *expectation* of this difference (i.e., the bias) is $O(n^{-1})$.

The unbiased estimates of the remaining two terms are

$$\widehat{B} := \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq}$$

and

$$\widehat{D} := \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij}l_{qr}.$$

While these expressions are unbiased, they are at first sight much more expensive to compute than the respective biased estimates, with $\widehat{B}$ costing $O(n^3)$ and $\widehat{D}$ costing $O(n^4)$. It is possible, however, to obtain these unbiased estimates in $O(n^2)$, i.e., the same cost as the biased estimates, as shown by [7, Theorem 2]. First, we note that diagonal entries of the kernel matrices K and L never appear in the sums, hence we immediately replace these matrices with $\widetilde{K}$ and $\widetilde{L}$ having the diagonal terms set to zero. The term $\widehat{A}$ can be written concisely in matrix form as

$$\widehat{A} = \frac{1}{(n)_2}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} = \frac{1}{(n)_2}\mathrm{trace}\left(\widetilde{K}\widetilde{L}\right),$$

where $\odot$ is the entrywise matrix product and $(A)_{++}$ is the sum of all the entries in A. Looking next at the term $\widehat{B}$, and defining as 1_n the $n\times 1$ vector of ones, we have

$$\begin{aligned}\widehat{B} = \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq} &= \frac{1}{(n)_3}\left[\sum_{i,j=1}^{n}\sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} - \sum_{i=1}^{n}\sum_{q\neq i}^{n} k_{iq}l_{iq}\right] \\ &= \frac{1}{(n)_3}1_n^\top\begin{bmatrix} \sum_{j=2}^{n} k_{1q}l_{q1} & \cdots & \sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} & \cdots \\ & \ddots & \vdots & \end{bmatrix}1_n \\ &\quad - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} \\ &= \frac{1}{(n)_3}1_n^\top\tilde{K}\tilde{L}1_n - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++}.\end{aligned}$$

The first expression in the final line can be computed in time $O(n^2)$, as long as the matrix-vector products are taken first. Finally, looking at the fourth term,[6]

$$\begin{aligned}
\widehat{D} &= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} = \frac{1}{(n)_4} \Bigg[\sum_{(i,j)\in \mathbf{i}_2^n} \sum_{(q,r)\in \mathbf{i}_2^n} k_{ij} l_{qr} \\
&\quad - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{ir}}_{q=i} - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{jr}}_{q=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(q=i,r=j)\equiv(q=j,r=i)} \\
&\quad - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq}}_{r=i} - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{jq}}_{r=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(r=i,q=j)\equiv(r=j,q=i)} \Bigg] \\
&= \frac{1}{(n)_4} \left[\left(\sum_{i=1}^n \sum_{j\neq i}^n k_{ij} \right) \left(\sum_{i=1}^n \sum_{j\neq i}^n l_{ij} \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right] \\
&= \frac{1}{(n)_4} \left[\left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right],
\end{aligned}$$

which can also be computed in $O(n^2)$. We now establish the net contribution of each term:

$$\begin{aligned}
\left(\widetilde{K} \odot \widetilde{L} \right)_{++} : & \frac{1}{(n)_2} + \frac{2}{(n)_3} + \frac{2}{(n)_4} \\
&= \frac{(n-2)(n-3) + (2n-6) + 2}{(n)_4} \\
&= \frac{(n-2)(n-1)}{(n)_4}
\end{aligned}$$

and

$$\begin{aligned}
1_n^\top \tilde{K} \tilde{L} 1_n \ : & \frac{-2}{(n)_3} - \frac{4}{(n)_4} \\
&= \frac{-2(n-3) - 4}{(n)_4} = \frac{-2(n-1)}{(n)_4}.
\end{aligned}$$

Thus, we have our empirical unbiased HSIC expression,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \frac{1}{n(n-3)} \left[\left(\widetilde{K} \odot \widetilde{L} \right)_{++} - \frac{2}{(n-2)} 1_n^\top \tilde{K} \tilde{L} 1_n + \frac{1}{(n-1)(n-2)} \left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) \right]$$

[6]The equivalences $\equiv$ in the first line below indicate that both index matching constraints amount to the same thing, hence these terms appear only once.

5 HSIC for feature selection

As we saw in the previous section, a biased estimate for the centred HSIC can be written

$$\text{HSIC} := \frac{1}{n^2}\text{trace}(KHLH).$$

Consider the case where we wish to find a subset of features that maximizes HSIC with respect to some set of labels. Assume we have a sample $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, and binary class labels. We choose a particular form for the class labels: $y_i \in \left\{n_+^{-1}, -n_-^{-1}\right\}$, where n_+ is the number of positive labels and n_- is the number of negative labels.

We denote by $x_i[\ell]$ the ℓth coordinate of x_i, and write

$$x[\ell] := \left[\begin{array}{ccc} x_1[\ell] & \ldots & x_n[\ell] \end{array}\right]^\top$$

the column vector of the ℓth coordinate of *all* samples. If we use a linear kernel on the x_i, then

$$K_{i,j} = x_i^\top x_j = \sum_{\ell=1}^{d} x_i[\ell]x_j[\ell].$$

It follows we can write the kernel as the sum of kernels on individual dimensions,

$$K = \sum_{\ell=1}^{d} K_\ell,$$

where $K_\ell := x[\ell]x[\ell]^\top$. In this case, HSIC is the sum of HSIC values for each such kernel,

$$\text{HSIC} := \frac{1}{n^2}\sum_{\ell=1}^{d}\text{trace}(K_\ell HLH).$$

What happens when we choose a linear kernel on the labels? Assuming the classes are grouped together,

$$L = yy^\top = \left[\begin{array}{cc} n_+^{-2}\mathbf{I} & -n_+n_-\mathbf{I} \\ -n_+n_-\mathbf{I} & n_-^{-2}\mathbf{I} \end{array}\right],$$

where y is the vector of all class labels. Note further than

$$\sum_{i=1}^{n} y_i = 0,$$

and hence $HLH = L$. Finally, using $\mathrm{trace}(AB) = \mathrm{trace}(BA)$,

$$\begin{aligned}
\mathrm{HSIC} &= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(K_\ell L) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(x[\ell]x[\ell]^\top yy^\top) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\left(\frac{1}{n_+}\sum_{i=1}^{n_+}x_i[l] - \frac{1}{n_-}\sum_{i=n_++1}^{n}x_i[l]\right)^2
\end{aligned}$$

6 Acknowledgments

Thanks to Aaditya Ramdas, Wittawat Jitkrittum, and Dino Sejdinovic for corrections and improvements to these notes.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

[3] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference*, pages 63–78, 2005.

[4] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[5] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. Version 20081110.

[6] M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis.* Academic Press, San Diego, 1980.

[7] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[8] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Information Science and Statistics. Springer, 2008.

[9] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proc. Annual Conf. Computational Learning Theory*, 2004.

Notes on mean embeddings and covariance operators

Arthur Gretton

January 11, 2020

1 Introduction

This note contains more detailed proofs of certain results in the lecture notes on mean embeddings and covariance operators. The notes are not as complete as for lectures 1 and 2, but cover only the trickier concepts. Please let me know if there are any further parts you'd like clarified, and I'll add them to the note.

2 Mean embeddings

2.1 Proof that the mean embedding exists via Riesz

For finite dimensional feature spaces, we can define expectations in terms of inner products.

$$\phi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

Then

$$f(x) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} = \langle f, \phi(x) \rangle_{\mathcal{F}} .$$

Consider random variable $\mathsf{x} \sim \mathbf{P}$

$$\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) = \mathbf{E}_{\mathbf{P}} \left(\begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathsf{x} \\ \mathsf{x}^2 \end{bmatrix} \right) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathbf{E}_{\mathbf{P}} \mathsf{x} \\ \mathbf{E}_{\mathbf{P}} (\mathsf{x}^2) \end{bmatrix} =: \begin{bmatrix} a \\ b \end{bmatrix}^\top \mu_{\mathbf{P}}.$$

Does this reasoning translate to infinite dimensions?

Definition 1 (Bounded operator)**.** A linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$Af \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

We prove via Riesz that the mean embedding exists, and that it takes the form of the expectation of the canonical map.

Theorem 2. *[Riesz representation]In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,*

$$Af = \langle f, g_A \rangle_{\mathcal{F}}$$

Now we establish the existence of the mean embedding.

Lemma 3 (Existence of mean embedding)**.** *If $\mathbf{E}_{\mathsf{P}} \sqrt{k(\mathsf{x}, \mathsf{x})} < \infty$ then $\mu_{\mathsf{P}} \in \mathcal{F}$.*

Proof. The linear operator $T_{\mathsf{P}} f := \mathbf{E}_{\mathsf{P}} f(\mathsf{x})$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_{\mathsf{P}} f| = |\mathbf{E}_{\mathsf{P}} f(\mathsf{x})| \underset{(a)}{\leq} \mathbf{E}_{\mathsf{P}} |f(\mathsf{x})| = \mathbf{E}_{\mathsf{P}} |\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}}| \leq \mathbf{E}_{\mathsf{P}} \left(\sqrt{k(\mathsf{x}, \mathsf{x})} \|f\|_{\mathcal{F}} \right),$$

where in (a) we use Jensen's inequality. Hence by the Riesz representer theorem [6, Theorem II.4], there exists a $\mu_{\mathsf{P}} \in \mathcal{F}$ such that $T_{\mathsf{P}} f = \langle f, \mu_{\mathsf{P}} \rangle_{\mathcal{F}}$. □

If we set $f = \phi(x) = k(x, \cdot)$, we obtain $\mu_{\mathsf{P}}(x) = \langle \mu_p, k(x, \cdot) \rangle = \mathbf{E}_{\mathsf{P}} k(x, \mathsf{x})$: in other words, the mean embedding of the distribution P is the expecation under P of the canonical feature map.

2.2 Proof that MMD injective for universal kernel

First, it is clear that $\mathsf{P} = \mathsf{Q}$ implies MMD $\{\mathsf{P}, \mathsf{Q}; F\}$ is zero. We now prove the converse. By the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ there exists a $g \in \mathcal{F}$ such that

$$\|f - g\|_{\infty} \leq \epsilon.$$

We will need [2, Lemma 9.3.2]:

Lemma 4. *Let $(\mathcal{X}, d)$ be a metric space, and let P, Q be two Borel probability measures defined on $\mathcal{X}$, where we define the random variables $\mathsf{x} \sim \mathsf{P}$ and $\mathsf{y} \sim \mathsf{Q}$. Then $\mathsf{P} = \mathsf{Q}$ if and only if $\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) = \mathbf{E}_{\mathsf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.*

We now use these two results to formulate a proof. We begin with the expansion

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq |\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| + |\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y})| + |\mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})|.$$

The first and third terms satisfy

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| \leq \mathbf{E}_{\mathsf{P}} |f(\mathsf{x}) - g(\mathsf{x})| \leq \epsilon.$$

Next, write

$$\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) = \langle g, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle = 0,$$

since MMD $\{\mathsf{P}, \mathsf{Q}; F\} = 0$ implies $\mu_{\mathsf{P}} = \mu_{\mathsf{Q}}$. Hence

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq 2\epsilon$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies $\mathsf{P} = \mathsf{Q}$ by Lemma 4.

3 Covariance operators

One of the most important and widely used tools in RKHS theory is the covariance operator: this is an infinite dimensional analog to the covariance matrix. This forms the backbone of kernel PCA, the kernel Fisher discriminant, kernel partial least squares, the kernel canonical correlation, and so on.

In this note, we describe the Hilbert space of Hilbert-Schmidt operators. We then introduce the covariance operator, demonstrate it is Hilbert-Schmidt, and express it in terms of kernel functions.

3.1 Hilbert-Schmidt operators

This discussion is based on [9, Section 2.1] and [8, Section A.5.2].

Let $\mathcal{F}$ and $\mathcal{G}$ be separable Hilbert spaces. Define $(e_i)_{i\in I}$ to be an orthonormal basis for $\mathcal{F}$, and $(f_j)_{j\in J}$ to be an orthonormal basis for $\mathcal{G}$. The index sets I, J are assumed to be either finite our countably infinite.[1] Define two compact linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$. Define the Hilbert-Schmidt norm of the operators L, M to be

$$\begin{aligned} \|L\|_{\mathrm{HS}}^2 &= \sum_{j\in J} \|Lf_j\|_{\mathcal{F}}^2 \\ &= \sum_{i\in I}\sum_{j\in J} \left|\langle Lf_j, e_i\rangle_{\mathcal{F}}\right|^2, \end{aligned} \tag{3.1}$$

where we use Parseval's indentity on each of the norms in the first sum. The operator L is Hilbert-Schmidt when this norm is finite.

The Hilbert-Schmidt operators mapping from $\mathcal{G}$ to $\mathcal{F}$ form a Hilbert space, written $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{j\in J} \langle Lf_j, Mf_j\rangle_{\mathcal{F}}, \tag{3.2}$$

which is independent of the orthonormal basis chosen. It is clear the norm (3.1) is recovered from this inner product. Another form for this inner product is

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{i\in I}\sum_{j\in J} \langle Lf_j, e_i\rangle_{\mathcal{F}} \langle Mf_j, e_i\rangle_{\mathcal{F}}. \tag{3.3}$$

Proof. Since any element of $\mathcal{F}$ can be expanded in terms of its orthonormal basis, we have that this holds in the specific case of the mapping of f_j by L or M,

$$Lf_j = \sum_{i\in I} \alpha_i^{(j)} e_i \qquad Mf_j = \sum_{i'\in I} \beta_{i'}^{(j)} e_{i'}. \tag{3.4}$$

[1] Recall that a Hilbert space has a countable orthonormal basis if and only if it is separable: that is, it has a countable dense subset [6, p. 47].

Substituting these into (3.2), we obtain

$$\begin{aligned}\langle L, M\rangle_{\mathrm{HS}} &= \sum_{j\in J}\left\langle \sum_{i\in I}\alpha_i^{(j)} e_i, \sum_{i'\in I}\beta_{i'}^{(j)} e_{i'}\right\rangle_{\mathcal{F}} \\ &= \sum_{i\in I}\sum_{j\in J}\alpha_i^{(j)}\beta_i^{(j)}.\end{aligned}$$

We obtain the identical result when we substitute (3.4) into (3.3). □

3.2 Rank-one operators, tensor product space

Given $b \in \mathcal{G}$ and $a \in \mathcal{F}$, we define the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$,

$$(b \otimes a) f \;\mapsto\; \langle f, a\rangle_{\mathcal{F}}\, b. \tag{3.5}$$

This is a generalization of the standard outer product in linear algebra, $(ba^\top)f = (a^\top f)b$, if all three of a, b, f were vectors. First, is this operator Hilbert-Schmidt? We compute its norm according to (3.1),

$$\begin{aligned}\|a \otimes b\|_{\mathrm{HS}}^2 &= \sum_{j\in J}\|(a\otimes b) f_j\|_{\mathcal{F}}^2 \\ &= \sum_{j\in J}\left\|a\,\langle b, f_j\rangle_{\mathcal{G}}\right\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j\in J}\left|\langle b, f_j\rangle_{\mathcal{G}}\right|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2,\end{aligned} \tag{3.6}$$

where we use Parseval's identity. Thus, the operator is Hilbert-Schmidt.

Given a second Hilbert-Schmidt operator $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, we have the result:

$$\langle L, a\otimes b\rangle_{\mathrm{HS}} = \langle a, Lb\rangle_{\mathcal{F}} \tag{3.7}$$

A particular instance of this result is

$$\langle u\otimes v, a\otimes b\rangle_{\mathrm{HS}} = \langle u, a\rangle_{\mathcal{F}}\,\langle b, v\rangle_{\mathcal{G}}\,. \tag{3.8}$$

Proof. The key result we use is the expansion of b in terms of the orthonormal basis, $b = \sum_{j\in J}\langle b, f_j\rangle_{\mathcal{G}}\, f_j$. Then

$$\begin{aligned}\langle a, Lb\rangle &= \left\langle a, L\left(\sum_j \langle b, f_j\rangle_{\mathcal{G}} f_j\right)\right\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}}\,\langle a, Lf_j\rangle_{\mathcal{F}}\end{aligned}$$

and

$$\begin{aligned}\langle a \otimes b, L\rangle_{\mathrm{HS}} &= \sum_j \langle Lf_j, (a \otimes b) f_j\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}} \langle Lf_j, a\rangle_{\mathcal{F}}.\end{aligned}$$

To show (3.8), we simply substitute $u \otimes v$ for L above, and then apply the definition (3.5),

$$\begin{aligned}\langle u \otimes v, a \otimes b\rangle_{\mathrm{HS}} \langle a, (u \otimes v)\, b\rangle_{\mathcal{F}} \\ = \langle u, a\rangle_{\mathcal{F}} \langle b, v\rangle_{\mathcal{G}}\end{aligned}$$

□

3.3 Cross-covariance operator

In this section, we define the cross-covariance operator, in the case where $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert spaces with respective kernels k and l, and feature maps ϕ and ψ. This is a generalization of the cross-covariance matrix to infinite dimensional feature spaces. The results we want are feature space analogues to:

$$\widetilde{C}_{XY} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right) \qquad f^\top \widetilde{C}_{XY} g = \mathbf{E}_{\mathsf{x}\mathsf{y}}\left[\left(f^\top \mathsf{x}\right)\left(g^\top \mathsf{y}\right)\right],$$

where we use the notation $\widetilde{C}_{XY}$ to denote a covariance operator without centering. The corresponding centered covariance is

$$C_{XY} := \widetilde{C}_{XY} - \mu_X \mu_Y^\top,$$

where $\mu_X := \mathbf{E}(\mathsf{x})$ and $\mu_Y := \mathbf{E}(\mathsf{y})$We now describe how we can get these results in feature space.

The cross product $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ is a random variable in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$: use the result in [9, p. 265] that for all $A \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, the linear form $\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}}$ is measurable. For the expectation of this random variable to exist (and to be an element of $\mathrm{HS}(\mathcal{G}, \mathcal{F})$), we require the expected norm of $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ to be bounded: in other words, $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$. Given the expectation exists, and writing it $\widetilde{C}_{XY}$, then this expectation is the unique element satisfying

$$\left\langle \widetilde{C}_{XY}, A\right\rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \tag{3.9}$$

Proof. The operator

$$\begin{aligned} T_{\mathsf{x}\mathsf{y}} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \mathbb{R} \\ A &\mapsto \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \end{aligned}$$

is bounded when $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$, since by applying first Jensen's inequality, then Cauchy-Schwarz,

$$
\begin{aligned}
|\mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| &\leq \mathbf{E}_{\mathsf{x},\mathsf{y}} |\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| \\
&\leq \|A\|_{\mathrm{HS}} \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right).
\end{aligned}
$$

Thus by the Riesz representer theorem (Theorem (2)), the covariance operator (3.9) exists. We can make a further simplification to the condition: substituting (3.6), we get the requirement

$$
\begin{aligned}
\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x})\|_{\mathcal{F}} \|\phi(\mathsf{y})\|_{\mathcal{G}} \right) \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\sqrt{k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y})} \right) < \infty.
\end{aligned}
$$

We could also use the weaker condition $\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y}) \right)$, which is implied from the above by Jensen's inequality. □

We now use the particular element $f \otimes g$. Combining (3.7) and (3.9), we have the result

$$
\begin{aligned}
\left\langle f, \widetilde{C}_{XY} g \right\rangle_{\mathcal{F}} &= \left\langle \widetilde{C}_{XY}, f \otimes g \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), f \otimes g \rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}} \langle g, \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[f(\mathsf{x}) g(\mathsf{y}) \right] = \mathrm{cov}(f, g).
\end{aligned}
$$

What does this operator look like? To see this, we apply it to $k(x, \cdot) l(y, \cdot)$ (just as we plotted the mean embedding by evaluating it on $k(x, \cdot)$).

We are given an i.i.d. sample from $\mathbf{P} = \mathbf{P}_{\mathsf{x}} \mathbf{P}_{\mathsf{y}}$, written $\mathbf{z} := ((x_1, y_1) \dots (x_n, y_n))$. Write the empirical

$$
\widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y,
$$

where we have now included the *centering terms* $\hat{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$. With some algebra, this can be written

$$
\widehat{C}_{XY} = \frac{1}{n} X H Y^\top,
$$

where $H = I_n - n^{-1} \mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix of ones, and

$$
X = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} \psi(y_1) & \dots & \psi(y_n) \end{bmatrix}.
$$

Define the kernel matrices

$$
K_{ij} = \left(X^\top X \right)_{ij} = k(x_i, x_j) \qquad L_{ij} = l(y_i, y_j),
$$

and the kernel matrices between centred variables,

$$
\widetilde{K} = HKH \qquad \widetilde{L} = HLH
$$

(exercise: prove that the above are kernel matrices for the variables centred in feature space).

4 Using the covariance operator to detect dependence

There are two measures of dependence we consider: the constrained covariance (COCO), which is the largest singular value of the covariance operator, and the Hilbert-Schmidt Independence Criterion, which is its Hilbert-Schmidt norm.

4.1 Empirical COCO and proof

We now derive the functions satisying

$$\begin{aligned} \text{maximize} \quad & \left\langle g, \widehat{C}_{XY} f \right\rangle_{\mathcal{G}} \\ \text{subject to} \quad & \|f\|_{\mathcal{F}} = 1 \qquad (4.1) \\ & \|g\|_{\mathcal{G}} = 1 \qquad (4.2) \end{aligned}$$

We assume that

$$f = \sum_{i=1}^{n} \alpha_i \left[\phi(x_i) - \hat{\mu}_x\right] = XH\alpha \qquad g = \sum_{j=1}^{n} \beta_i \left[\psi(y_i) - \hat{\mu}_y\right] = YH\beta,$$

where

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \qquad \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i).$$

The associated Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = -f^\top \widehat{C}_{XY} g + \frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right),$$

where we have negated the covariance to make it a minimization problem (for consistency with the optimisation lecture later in the course), and we divide the Lagrange multipliers by 2 to simplify the discussion later . We now write this in terms of α and β:

$$\begin{aligned} f^\top \widehat{C}_{XY} g &= \frac{1}{n}\alpha^\top H X^\top \left(XHY^\top\right) YH\beta \\ &= \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta, \end{aligned}$$

where we note that $H = HH$. Similarly

$$\|f\|_{\mathcal{F}}^2 = \alpha^\top HXX^\top H\alpha = \alpha^\top \widetilde{K}\alpha.$$

Substituting these into the Lagrangian, we get a new optimization in terms of α and β,

$$\mathcal{L}(\alpha, \beta, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right). \qquad (4.3)$$

We must minimise this wrt the primal variables α, β. Differentiating wrt α and β and setting the resulting expressions to zero,[2] we obtain

$$-\frac{1}{n}\widetilde{K}\widetilde{L}\beta + \lambda\widetilde{K}\alpha = 0 \tag{4.4}$$

$$-\frac{1}{n}\widetilde{L}\widetilde{K}\alpha + \gamma\widetilde{L}\beta = 0 \tag{4.5}$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta = \lambda\alpha^\top\widetilde{K}\alpha$$
$$\frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta$$

Subtracting the first expression from the second, we get

$$\lambda\alpha^\top\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta.$$

Thus for $\lambda \neq 0$ and $\gamma \neq 0$, we conclude that $\lambda = \gamma$. Making this replacement in (4.4) and (4.5), we must find the largest[3] γ that solves the following expression wrt α, β:

$$\begin{bmatrix} 0 & \frac{1}{n}\widetilde{K}\widetilde{L} \\ \frac{1}{n}\widetilde{L}\widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{4.6}$$

This is a generalized eigenvalue problem, and can be solved straightforwardly in Matlab. The maximum eigenvalue is indeed COCO: at the solution, $\alpha^\top\widetilde{K}\alpha = 1$ and $\beta^\top\widetilde{L}\beta = 1$, hence the two norm terms in the Lagrangian (4.3) vanish.[4]

[2] We use [5, eqs. (61) and (73)]

$$\frac{\partial a^\top U a}{\partial a} = (U + U^\top)a, \qquad \frac{\partial v^\top a}{\partial a} = \frac{\partial a^\top v}{\partial a} = v.$$

[3] Given $\lambda = \gamma$, the system of equations (4.4) and (4.5) becomes:

$$\frac{1}{n}\widetilde{K}\widetilde{L}\beta = \gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = \gamma\widetilde{L}\beta$$

However we also get a valid solution by switching $\check{\beta} := -\beta$,

$$\frac{1}{n}\widetilde{K}\widetilde{L}\check{\beta} = -\gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = -\gamma\widetilde{L}\check{\beta}$$

In other words, the solutions γ of the generalised eigenvalue problem (4.6) come in pairs $\pm\gamma$, depending on the relative sign of α and β.

[4] For a more roundabout way of reaching the same conclusion: pre-multiply (4.6) by $[\alpha^\top \beta^\top]$ to get the system of equations

$$\begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta \\ \frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha \end{bmatrix} = \gamma \begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\alpha \\ \frac{1}{n}\beta^\top\widetilde{L}\beta \end{bmatrix} = \gamma \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where in the final line we substitute the constraints from (4.1).

4.2 The Hilbert-Schmidt Independence Criterion

4.2.1 Population expression

What is the Hilbert-Schmidt norm of the covariance operator?[5] Consider the centered, squared norm of the RKHS covariance operator,

$$\begin{aligned} HSIC^2(\mathcal{F}, \mathcal{G}, P_{xy}) &= \|\widetilde{C}_{XY} - \mu_X \otimes \mu_Y\|^2_{\text{HS}} \\ &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} - 2 \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}}, \end{aligned}$$

where $\widetilde{C}_{XY}$ is the uncentered covariance operator defined in (3.9). There are three terms in the expansion.

To obtain the first term, we apply (3.9) twice, denoting by $(\mathsf{x}', \mathsf{y}')$ an independent copy of the pair of variables (x, y),

$$\begin{aligned} \|\widetilde{C}_{XY}\|^2_{\text{HS}} &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \phi(\mathsf{x}') \otimes \psi(\mathsf{y}') \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}), [\phi(\mathsf{x}') \otimes \psi(\mathsf{y}')] \psi(\mathsf{y}) \rangle_{\mathcal{F}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left[\langle \phi(\mathsf{x}), \phi(\mathsf{x}') \rangle_{\mathcal{F}} \langle \psi(\mathsf{y}'), \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}') \\ &=: A \end{aligned}$$

Similar reasoning can be used to show

$$\begin{aligned} \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} &= \langle \mu_X, \mu_X \rangle_{\mathcal{F}} \langle \mu_Y, \mu_Y \rangle_{\mathcal{G}} \\ &= \mathbf{E}_{\mathsf{x}\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \\ &=: D, \end{aligned}$$

and for the cross-terms,

$$\begin{aligned} \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}} &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \mu_X \otimes \mu_Y \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\langle \phi(\mathsf{x}), \mu_X \rangle_{\mathcal{F}} \langle \phi(\mathsf{y}), \mu_Y \rangle_{\mathcal{G}} \right) \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\mathbf{E}_{\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \right) \\ &=: B. \end{aligned}$$

[5] Other norms of the operator may also be used in determining dependence, e.g. the spectral norm from the previous section. Another statistic on the kernel spectrum is the Kernel Mutual Information, which is an upper bound on the true mutual information near independence, but is otherwise difficult to interpret [4]. One can also define independence statistics on the correlation operator [1], which may be better behaved for small sample sizes, although the asymptotic behavior is harder to analyze.

4.2.2 Biased estimate

A biased estimate of HSIC was given in [3]. We observe a sample $Z := \{(x_1, y_1), \dots (x_n, y_n)\}$ drawn independently and identically from P_{xy}, we wish to obtain empirical expressions for HSIC,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \widehat{A} - 2\widehat{B} + \widehat{D}.$$

A direct approach would be to replace the population uncentred covariance operator $\widetilde{C}_{XY}$ with an empirical counterpart,

$$\check{C}_{XY} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i),$$

and the population mean embeddings with their respective empirical estimates,

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \qquad \hat{\mu}_y = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i),$$

however the resulting estimates are biased (we will show the amount of bias in the next section). The first term is

$$\begin{aligned} \widehat{A}_b = \left\| \check{C}_{XY} \right\|^2 &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} l_{ij} = \frac{1}{n^2} \text{tr}(KL), \end{aligned}$$

we use the shorthand $k_{ij} = k(x_i, x_j)$, and the subscript b to denote a biased estimate. The expression is not computationally efficient, and is written this way for later use - in practice, we would never take the matrix product if the intent was then to compute the trace. Next,

$$\begin{aligned} \widehat{B}_b = \left\langle \check{C}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y \right\rangle &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \otimes \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(x_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \right\rangle_{F} \left\langle \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\mathcal{G}} \\ &= \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{q=1}^{n} k_{ij} l_{iq} \\ &= \frac{1}{n^3} 1_n^\top KL 1_n = \frac{1}{n^3} 1_n^\top LK 1_n \end{aligned}$$

(we will use both forms to get our final biased estimate of HSIC), and

$$\begin{aligned}\widehat{D}_b = \langle \hat{\mu}_X \otimes \hat{\mu}_Y, \hat{\mu}_X \otimes \hat{\mu}_Y \rangle &= \left\langle \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right), \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^4}\left(\sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j)\right)\left(\sum_{i=1}^n \sum_{j=1}^n l(y_i, y_j)\right) \\ &= \frac{1}{n^4}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\end{aligned}$$

We now combine these terms, to obtain the biased estimate

$$\begin{aligned}HSIC_b^2(\mathcal{F}, \mathcal{G}, Z) &= \frac{1}{n^2}\left(\text{tr}(KL) - \frac{2}{n} 1_n^\top KL 1_n + \frac{1}{n^2}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\right) \\ &= \frac{1}{n^2}\left[\text{tr}(KL) - \frac{1}{n}\text{tr}\left(1_n 1_n^\top KL\right) - \frac{1}{n}\text{tr}\left(K 1_n 1_n^\top L\right) + \frac{1}{n^2}\text{tr}\left(1_n 1_n^\top K 1_n 1_n^\top L\right)\right] \\ &= \frac{1}{n^2}\text{tr}\left[\left(I - \frac{1}{n} 1_n 1_n^\top\right) K \left(I - \frac{1}{n} 1_n 1_n^\top\right) L\right] \\ &= \frac{1}{n^2}\text{tr}(KHLH)\end{aligned}$$

where we define

$$H := I - \frac{1}{n} 1_n 1_n^\top$$

as a centering matrix (when pre-multiplied by a matrix it centers the rows; when post-multiplied, it centers the columns).

4.2.3 Unbiased estimate

An unbiased estimate of $A := \|\widetilde{C}_{XY}\|_{\text{HS}}^2$ is

$$\widehat{A} := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k_{ij} l_{ij} = \frac{1}{(n)_2} \sum_{(i,j) \in \mathbf{i}_2^n} k_{ij} l_{ij},$$

where $\mathbf{i}_p^n$ is the set of all p-tuples drawn from $\{1, \ldots, n\}$, and

$$(n)_p = \frac{n!}{(n-p)!} = \frac{1}{n(n-1)\ldots(n-p+1)}.$$

Note that $\mathbf{E}(\widehat{A}) = \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}')$, which is not true of the biased expression (which does not properly treat the independent copies x' of x and y' of

y). The difference between the biased and unbiased estimates is

$$\begin{aligned}\widehat{A}_b - \widehat{A} &= \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}l_{ij} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} k_{ii}l_{ii} + \left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right)\left(\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right) \\ &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} k_{ii}l_{ii} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right),\end{aligned}$$

thus the *expectation* of this difference (i.e., the bias) is $O(n^{-1})$.

The unbiased estimates of the remaining two terms are

$$\widehat{B} := \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq}$$

and

$$\widehat{D} := \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij}l_{qr}.$$

While these expressions are unbiased, they are at first sight much more expensive to compute than the respective biased estimates, with $\widehat{B}$ costing $O(n^3)$ and $\widehat{D}$ costing $O(n^4)$. It is possible, however, to obtain these unbiased estimates in $O(n^2)$, i.e., the same cost as the biased estimates, as shown by [7, Theorem 2]. First, we note that diagonal entries of the kernel matrices K and L never appear in the sums, hence we immediately replace these matrices with $\widetilde{K}$ and $\widetilde{L}$ having the diagonal terms set to zero. The term $\widehat{A}$ can be written concisely in matrix form as

$$\widehat{A} = \frac{1}{(n)_2}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} = \frac{1}{(n)_2}\mathrm{trace}\left(\widetilde{K}\widetilde{L}\right),$$

where $\odot$ is the entrywise matrix product and $(A)_{++}$ is the sum of all the entries in A. Looking next at the term $\widehat{B}$, and defining as 1_n the $n\times 1$ vector of ones, we have

$$\begin{aligned}\widehat{B} = \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq} &= \frac{1}{(n)_3}\left[\sum_{i,j=1}^{n}\sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} - \sum_{i=1}^{n}\sum_{q\neq i}^{n} k_{iq}l_{iq}\right] \\ &= \frac{1}{(n)_3}1_n^\top\begin{bmatrix} \sum_{j=2}^{n} k_{1q}l_{q1} & \cdots & \sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} & \cdots \\ & \ddots & \vdots & \end{bmatrix}1_n \\ &\quad - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} \\ &= \frac{1}{(n)_3}1_n^\top\tilde{K}\tilde{L}1_n - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++}.\end{aligned}$$

The first expression in the final line can be computed in time $O(n^2)$, as long as the matrix-vector products are taken first. Finally, looking at the fourth term,[6]

$$\begin{aligned}
\widehat{D} &= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} = \frac{1}{(n)_4} \Bigg[\sum_{(i,j)\in \mathbf{i}_2^n} \sum_{(q,r)\in \mathbf{i}_2^n} k_{ij} l_{qr} \\
&\quad - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{ir}}_{q=i} - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{jr}}_{q=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(q=i,r=j)\equiv(q=j,r=i)} \\
&\quad - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq}}_{r=i} - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{jq}}_{r=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(r=i,q=j)\equiv(r=j,q=i)} \Bigg] \\
&= \frac{1}{(n)_4} \left[\left(\sum_{i=1}^n \sum_{j\neq i}^n k_{ij} \right) \left(\sum_{i=1}^n \sum_{j\neq i}^n l_{ij} \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right] \\
&= \frac{1}{(n)_4} \left[\left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right],
\end{aligned}$$

which can also be computed in $O(n^2)$. We now establish the net contribution of each term:

$$\begin{aligned}
\left(\widetilde{K} \odot \widetilde{L} \right)_{++} : & \frac{1}{(n)_2} + \frac{2}{(n)_3} + \frac{2}{(n)_4} \\
&= \frac{(n-2)(n-3) + (2n-6) + 2}{(n)_4} \\
&= \frac{(n-2)(n-1)}{(n)_4}
\end{aligned}$$

and

$$\begin{aligned}
1_n^\top \tilde{K} \tilde{L} 1_n \ : & \frac{-2}{(n)_3} - \frac{4}{(n)_4} \\
&= \frac{-2(n-3) - 4}{(n)_4} = \frac{-2(n-1)}{(n)_4}.
\end{aligned}$$

Thus, we have our empirical unbiased HSIC expression,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \frac{1}{n(n-3)} \left[\left(\widetilde{K} \odot \widetilde{L} \right)_{++} - \frac{2}{(n-2)} 1_n^\top \tilde{K} \tilde{L} 1_n + \frac{1}{(n-1)(n-2)} \left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) \right]$$

[6]The equivalences $\equiv$ in the first line below indicate that both index matching constraints amount to the same thing, hence these terms appear only once.

5 HSIC for feature selection

As we saw in the previous section, a biased estimate for the centred HSIC can be written

$$\text{HSIC} := \frac{1}{n^2}\text{trace}(KHLH).$$

Consider the case where we wish to find a subset of features that maximizes HSIC with respect to some set of labels. Assume we have a sample $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, and binary class labels. We choose a particular form for the class labels: $y_i \in \left\{n_+^{-1}, -n_-^{-1}\right\}$, where n_+ is the number of positive labels and n_- is the number of negative labels.

We denote by $x_i[\ell]$ the ℓth coordinate of x_i, and write

$$x[\ell] := \left[\begin{array}{ccc} x_1[\ell] & \ldots & x_n[\ell] \end{array}\right]^\top$$

the column vector of the ℓth coordinate of *all* samples. If we use a linear kernel on the x_i, then

$$K_{i,j} = x_i^\top x_j = \sum_{\ell=1}^{d} x_i[\ell]x_j[\ell].$$

It follows we can write the kernel as the sum of kernels on individual dimensions,

$$K = \sum_{\ell=1}^{d} K_\ell,$$

where $K_\ell := x[\ell]x[\ell]^\top$. In this case, HSIC is the sum of HSIC values for each such kernel,

$$\text{HSIC} := \frac{1}{n^2}\sum_{\ell=1}^{d}\text{trace}(K_\ell HLH).$$

What happens when we choose a linear kernel on the labels? Assuming the classes are grouped together,

$$L = yy^\top = \left[\begin{array}{cc} n_+^{-2}\mathbf{I} & -n_+n_-\mathbf{I} \\ -n_+n_-\mathbf{I} & n_-^{-2}\mathbf{I} \end{array}\right],$$

where y is the vector of all class labels. Note further than

$$\sum_{i=1}^{n} y_i = 0,$$

and hence $HLH = L$. Finally, using $\mathrm{trace}(AB) = \mathrm{trace}(BA)$,

$$\begin{aligned}
\mathrm{HSIC} &= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(K_\ell L) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(x[\ell]x[\ell]^\top yy^\top) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\left(\frac{1}{n_+}\sum_{i=1}^{n_+}x_i[l] - \frac{1}{n_-}\sum_{i=n_++1}^{n}x_i[l]\right)^2
\end{aligned}$$

6 Acknowledgments

Thanks to Aaditya Ramdas, Wittawat Jitkrittum, and Dino Sejdinovic for corrections and improvements to these notes.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

[3] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference*, pages 63–78, 2005.

[4] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[5] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. Version 20081110.

[6] M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis.* Academic Press, San Diego, 1980.

[7] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[8] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Information Science and Statistics. Springer, 2008.

[9] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proc. Annual Conf. Computational Learning Theory*, 2004.

Support Vector machines

Arthur Gretton

November 30, 2018

1 Outline

- Review of convex optimization
- Support vector classification. C-SV and ν-SV machines

2 Review of convex optimization

This review covers the material from [1, Sections 5.1-5.5]. We begin with definitions of convex functions and convex sets.

2.1 Convex sets, convex functions

Definition 1 (Convex set)**.** A set C is convex if for all $x_1, x_2 \in C$ and any $0 \leq \theta \leq 1$ we have $\theta x_1 + (1-\theta)x_2 \in C$, i.e. every point on the line between x_1 and x_2 lies in C. See Figure 2.1.

In other words, every point in the set can be seen from any other point in the set, along a straight line that never leaves the set.

We next introduce the notion of a convex function.

Definition 2 (Convex function)**.** A function f is convex if its domain $\mathrm{dom} f$ is a convex set and if $\forall x, y \in \mathrm{dom} f$, and any $0 \leq \theta \leq 1$,

$$f\left(\theta x + (1-\theta)y\right) \leq \theta f(x) + (1-\theta)f(y).$$

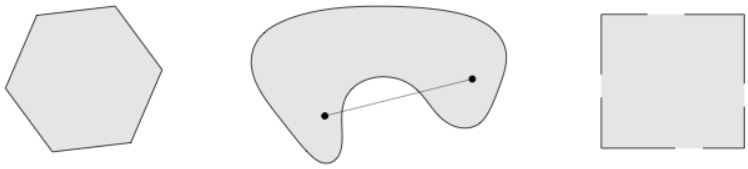

Figure 2.1: Examples of convex and non-convex sets (taken from [1, Fig. 2.2]). The first set is convex, the last two are not.

Figure 2.2: Convex function (taken from [1, Fig. 3.1])

The function is strictly convex if the inequality is strict for $x \neq y$. See Figure 2.2.

2.2 The Lagrangian

We now consider an optimization problem on $x \in \mathbb{R}^n$,

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, m \\ & h_i(x) = 0 & i = 1, \ldots p. \end{array} \tag{2.1}$$

We define by p^* the optimal value of (2.1), and by $\mathcal{D} := \bigcap_{i=0}^m \text{dom} f_i \cap \bigcap_{i=1}^p \text{dom} h_i$, where we require the domain[1] $\mathcal{D}$ to be nonempty.

The **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with problem (2.1) is written

$$L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x),$$

and has domain $\text{dom} L := \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The vectors λ and ν are called **lagrange multipliers** or **dual variables**. The **Lagrange dual function** (or just "dual function") is written

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu).$$

If this is unbounded below, then the dual is $-\infty$. The domain of g, $\text{dom}(g)$, is the set of values (λ, μ) such that $g > -\infty$. The dual function is a pointwise infimum of affine[2] functions of (λ, ν), hence it is concave in (λ, ν) [1, p. 83].

When[3] $\lambda \succeq 0$, then for all ν we have

$$g(\lambda, \nu) \leq p^*. \tag{2.2}$$

[1] The domain is the set on which a function is well defined. Eg the domain of $\log x$ is $\mathbb{R}^{++}$, the strictly positive real numbers [1, p. 639].

[2] A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is affine if it takes the form $f(x) = Ax + b$.

[3] The notation $a \succeq b$ for vectors a, b means that $a_i \geq b_i$ for all i.

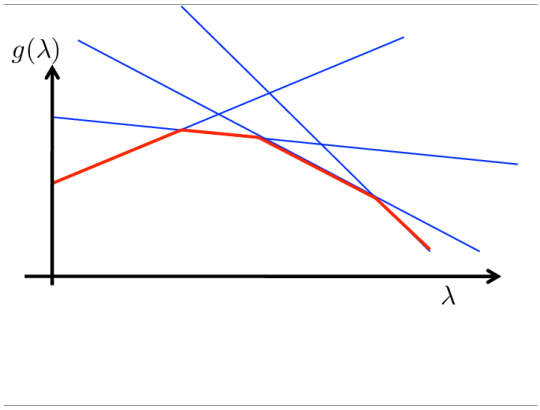

Figure 2.3: Example: Lagrangian with one inequality constraint, $L(x, \lambda) = f_0(x) + \lambda f_1(x)$, where x here can take one of four values for ease of illustration. The infimum of the resulting set of four affine functions is concave in λ.

See Figure (2.4) for an illustration on a toy problem with a single inequality constraint. A **dual feasible** pair (λ, ν) is a pair for which $\lambda \succeq 0$ and $(\lambda, \nu) \in \mathrm{dom}(g)$.

Proof. (of eq. (2.2)) Assume $\tilde{x}$ is feasible for the optimization, i.e. $f_i(\tilde{x}) \leq 0$, $h_i(\tilde{x}) = 0$, $\tilde{x} \in \mathcal{D}$, $\lambda \succeq 0$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq 0$$

and so

$$\begin{aligned} g(\lambda, \nu) &:= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \\ &\leq f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \\ &\leq f_0(\tilde{x}). \end{aligned}$$

This holds for every feasible $\tilde{x}$, hence (2.2) holds. □

We now give a lower bound interpretation. Idealy we would write the problem (2.1) as the unconstrained problem

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} I_-\left(f_i(x)\right) + \sum_{i=1}^{p} I_0\left(h_i(x)\right),$$

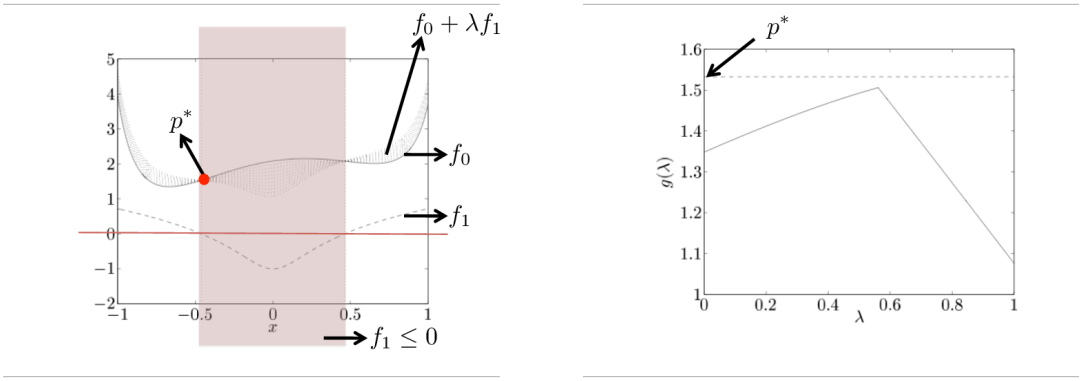

Figure 2.4: Illustration of the dual function for a simple problem with one inequality constraint (from [1, Figs. 5.1 and 5.2]). In the right hand plot, the dashed line corresponds to the optimum p^* of the original problem, and the solid line corresponds to the dual as a function of λ. Note that the dual as a function of λ is concave.

where

$$I_-(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$$

and $I_0(u)$ is the indicator of 0. This would then give an infinite penalization when a constraint is violated. Instead of these sharp indicator functions (which are hard to optimize), we replace the constraints with a set of soft linear constraints, as shown in Figure 2.5. It is now clear why λ must be positive for the inequality constraint: a negative λ would not yield a lower bound. Note also that as well as being penalized for $f_i > 0$, the linear lower bounds reward us for achieving $f_i < 0$.

2.3 The dual problem

The dual problem attempts to find the best lower bound $g(\lambda, \nu)$ on the optimal solution p^* of (2.1). This results in the **Lagrange dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{2.3}$$

We use **dual feasible** to describe (λ, ν) with $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$. The solutions to the dual problem are written (λ^*, ν^*), and are called **dual optimal**. Note that (2.3) is a convex optimization problem, since the function being maximized is concave and the constraint set is convex. We denote by d^* the optimal value of the dual problem. The property of **weak duality** always holds:

$$d^* \leq p^*.$$

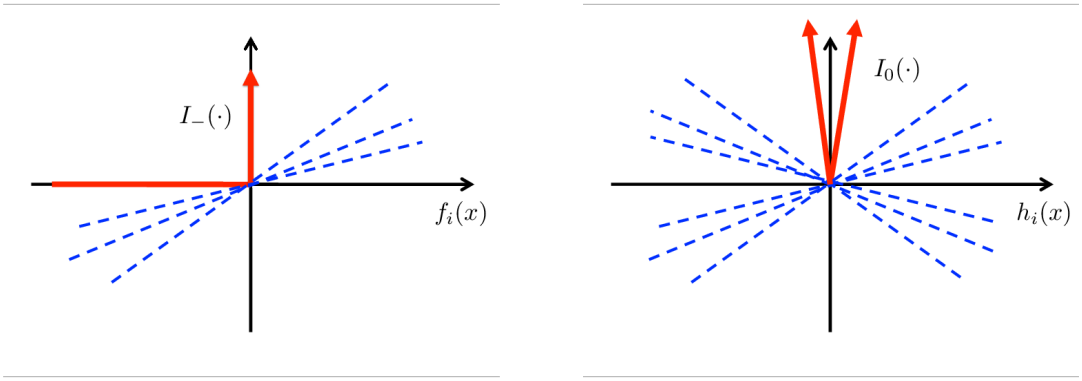

Figure 2.5: Linear lower bounds on indicator functions. Blue functions represent linear lower bounds for different slopes λ and ν, for the inequality and equality constraints, respectively.

The difference $p^* - d^*$ is called the **optimal duality gap**. If the duality gap is zero, then **strong duality** holds:

$$d^* = p^*.$$

Conditions under which strong duality holds are called **constraint qualifications**. As an important case: strong duality holds if the primal problem is convex,[4] i.e. of the form

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0 & i = 1, \ldots, n \\ & Ax = b & \end{array} \tag{2.4}$$

for convex $f_0, \ldots, f_m$, *and* if **Slater's condition** holds: there exists some *strictly* feasible point[5] $\tilde{x} \in \text{relint}(\mathcal{D})$ such that

$$f_i(\tilde{x}) < 0 \quad i = 1, \ldots, m \quad A\tilde{x} = b.$$

A weaker version of Slater's condition is sufficient for strong convexity when some of the constraint functions $f_1, \ldots, f_k$ are affine (note the inequality constraints are no longer strict):

$$f_i(\tilde{x}) \leq 0 \quad i = 1, \ldots, k \quad f_i(\tilde{x}) < 0 \quad i = k+1, \ldots, m \quad A\tilde{x} = b.$$

A proof of this result is given in [1, Section 5.3.2].

[4] Strong duality can also hold for non-convex problems: see e.g. [1, p. 229].

[5] We denote by $\text{relint}(\mathcal{D})$ the relative interior of the set $\mathcal{D}$. This looks like the interior of the set, but is non-empty even when the set is a subspace of a larger space. See [1, Section 2.1.3] for the formal defintion.

2.4 A saddle point/game characterization of weak and strong duality

In this section, we ignore equality constraints for ease of discussion. We write the solution to the primal problem as an optimization

$$\begin{aligned}\sup_{\lambda \succeq 0} L(x,\lambda) &= \sup_{\lambda \succeq 0} \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \right) \\ &= \begin{cases} f_0(x) & f_i(x) \leq 0,\ i = 1,\ldots,m \\ \infty & \text{otherwise.} \end{cases}\end{aligned}$$

In other words, we recover the primal problem when the inequality constraint holds, and get infinity otherwise. We can therefore write

$$p^* = \inf_x \sup_{\lambda \succeq 0} L(x,\lambda).$$

We already know

$$d^* = \sup_{\lambda \succeq 0} \inf_x L(x,\lambda).$$

Weak duality therefore corresponds to the **max-min inequality**:

$$\sup_{\lambda \succeq 0} \inf_x L(x,\lambda) \leq \inf_x \sup_{\lambda \succeq 0} L(x,\lambda). \tag{2.5}$$

which holds for general functions, and not just $L(x,\lambda)$. Strong duality occurs at a saddle point, and the inequality becomes an equality.

There is also a **game interpretation**: $L(x,\lambda)$ is a sum that must be paid by the person ajusting x to the person adjusting λ. On the right hand side of (2.5), player x plays first. Knowing that player 2 (λ) will maximize their return, player 1 (x) chooses their setting to give player 2 the worst possible options over all λ. The max-min inequality says that whoever plays second has the advantage.

2.5 Optimality conditions

If the primal is equal to the dual, we can make some interesting observations about the duality constraints. Denote by x^* the optimum solution of the original problem (the minimum of f_0 under its constraints), and by (λ^*, ν^*) the solutions to the dual. Then

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &\underset{\text{(a)}}{=} \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \\ &\underset{\text{(b)}}{\leq} f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\ &\leq f_0(x^*), \end{aligned}$$

where in (a) we use the definition of g, in (b) we use that $\inf_{x\in\mathcal{D}}$ of the expression in the parentheses is necessarily no greater than its value at x^*, and the last line we use that at (x^*, λ^*, ν^*) we have $\lambda^* \succeq 0$, $f_i(x^*) \leq 0$, and $h_i(x^*) = 0$. From this chain of reasoning, it follows that

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0, \tag{2.6}$$

which is the condition of **complementary slackness**. This means

$$\begin{aligned} \lambda_i^* > 0 \quad &\Longrightarrow \quad f_i(x^*) = 0, \\ f_i(x^*) < 0 \quad &\Longrightarrow \quad \lambda_i^* = 0. \end{aligned}$$

Consider now the case where the functions f_i, h_i are differentiable, and the duality gap is zero. Since x^* minimizes $L(x, \lambda^*, \nu^*)$, the derivative at x^* should be zero,

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0.$$

We now gather the various conditions for optimality we have discussed. The **KKT conditions** for the primal and dual variables (x, λ, ν) are

$$\begin{aligned} f_i(x) &\leq 0,\ i = 1, \ldots, m \\ h_i(x) &= 0,\ i = 1, \ldots, p \\ \lambda_i &\geq 0,\ i = 1, \ldots, m \\ \lambda_i f_i(x) &= 0,\ i = 1, \ldots, m \\ \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) &= 0 \end{aligned}$$

If a convex optimization problem with differentiable objective and constraint functions satisfies Slater's conditions, then the KKT conditions are necessary and sufficient for global optimality.

3 The representer theorem

This description comes from Lecture 8 of Peter Bartlett's course on Statistical Learning Theory.

We are given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$ (the setting could be regression or classification). We consider problems of a very general type: we want to find the function f in the RKHS $\mathcal{H}$ which satisfies

$$J(f^*) = \min_{f\in\mathcal{H}} J(f), \tag{3.1}$$

where

$$J(f) = L_y(f(x_1), \ldots, f(x_n)) + \Omega\left(\|f\|_{\mathcal{H}}^2\right),$$

Ω is non-decreasing, and y is the vector of y_i. Examples of loss functions might be

- Classification: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n \mathbb{I}_{y_i f(x_i) \leq 0}$ (the number of points for which the sign of y disagrees with that of the prediction $f(x)$),
- Regression: $L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^n (y_i - f(x_i))^2$, the sum of squared errors (eg. when $\Omega\left(\|f\|_{\mathcal{H}}^2\right) = \|f\|_{\mathcal{H}}^2$, we are back to the standard ridge regression setting).

The representer theorem states that a solution to 3.1 takes the form

$$f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot).$$

If Ω is strictly increasing, all solutions have this form.

Proof: We write as f_s the projection of f onto the subspace

$$\operatorname{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}, \tag{3.2}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$. Consider first the regularizer term. Since

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right),$$

so this term is minimized for $f = f_s$. Next, consider the individual terms $f(x_i)$ in the loss. These satisfy

$$f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(x_i, \cdot)\rangle_{\mathcal{H}} = \langle f_s, k(x_i, \cdot)\rangle_{\mathcal{H}},$$

so

$$L_y(f(x_1), \ldots, f(x_n)) = L_y(f_s(x_1), \ldots, f_s(x_n)).$$

Hence the loss $L(\ldots)$ only depends on the component of f in the subspace 3.2, and the regularizer $\Omega(\ldots)$ is minimized when $f = f_s$. If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum, otherwise this may be one of several minima.

4 Support vector classification

4.1 The linearly separable case

We consider problem of classifying two clouds of points, where there exists a hyperplane which linearly separates one cloud from the other without error.

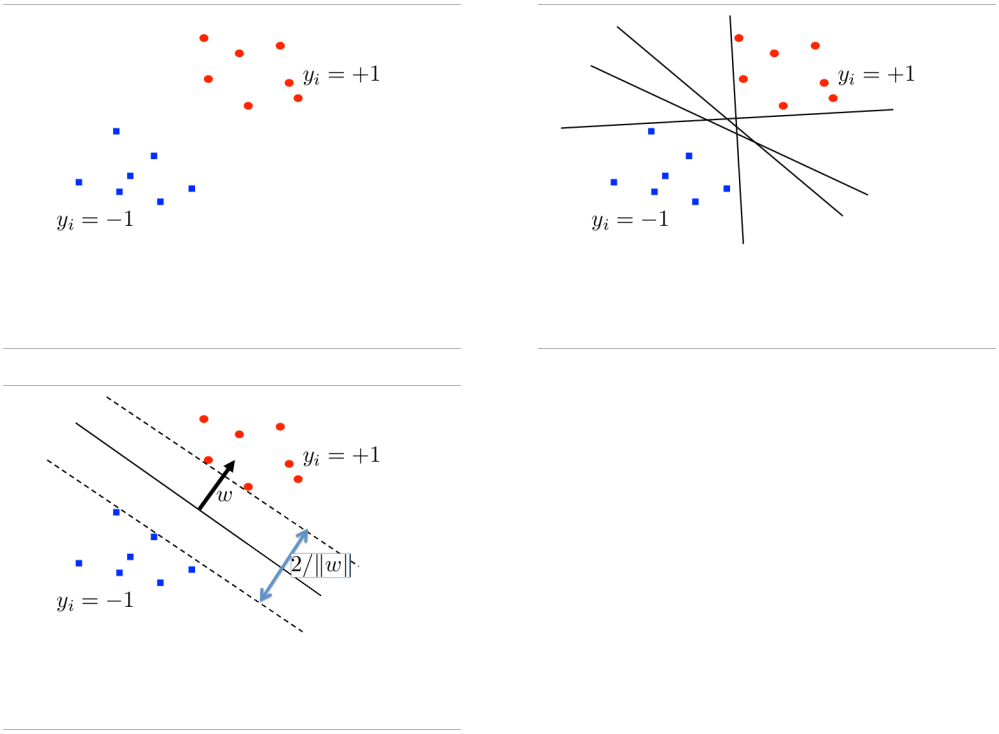

Figure 4.1: The linearly separable case. There are many linear separating hyperplanes, but only one max. margin separating hyperplane.

This is illustrated in Figure (4.1). As can be seen, there are infinitely many possible hyperplanes that solve this problem: the question is then: which one to choose? We choose the one which has the largest margin: it is the largest possible distance from both classes, and the *smallest* distance from each class to the separating hyperplane is called the margin.

This problem can be expressed as follows:[6]

$$\max_{w,b} (\text{margin}) = \max_{w,b} \left(\frac{2}{\|w\|} \right) \tag{4.2}$$

[6]It's easy to see why the equation below is the margin (the distance between the positive and negative classes): consider two points x_+ and x_- of opposite label, located on the margins. The width of the margin, d_m, is the difference $x_+ - x_-$ projected onto the unit vector in the direction w, or

$$d_m = (x_+ - x_-)^\top \frac{w}{\|w\|} \tag{4.1}$$

Subtracting the two equations in the constraints (4.3) from each other, we get

$$w^\top (x_+ - x_-) = 2.$$

Substituting this into (4.1) proves the result.

subject to

$$\begin{cases} \min\left(w^\top x_i + b\right) = 1 & i \,:\, y_i = +1, \\ \max\left(w^\top x_i + b\right) = -1 & i \,:\, y_i = -1. \end{cases} \tag{4.3}$$

The resulting classifier is

$$y = \text{sign}(w^\top x + b),$$

where sign takes value $+1$ for a positive argument, and -1 for a negative argument (its value at zero is not important, since for non-pathological cases we will not need to evaluate it there). We can rewrite to obtain

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{or} \quad \min_{w,b} \|w\|^2$$

subject to

$$y_i(w^\top x_i + b) \geq 1. \tag{4.4}$$

4.2 When no linear separator exists (or we want a larger margin)

If the classes are not linearly separable, we may wish to allow a certain number of errors in the classifier (points on the wrong side of the decision boudary). Ideally, we would optimise

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \mathbb{I}[y_i \left(w^\top x_i + b\right) < 0] \right),$$

where C controls the tradeoff between maximum margin and loss, and $\mathbb{I}(A) = 1$ if A holds true, and 0 otherwise (the factor of $1/2$ is to simplify the algebra later, and is not important: we can adjust C accordingly). This is a combinatorical optimization problem, which would be very expensive to solve. Instead, we replace the indicator function with a convex upper bound,

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

We use the hinge loss,

$$\theta(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise.} \end{cases}$$

although obviously other choices are possible (e.g. a quadratic upper bound). See Figure 4.2.

Substituting in the hinge loss, we get

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \theta\left(y_i \left(w^\top x_i + b\right)\right) \right).$$

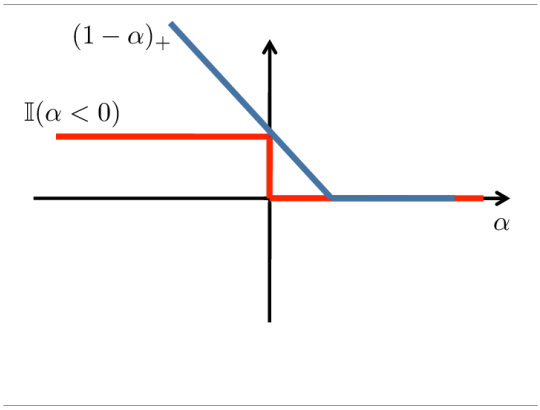

Figure 4.2: The hinge loss is an upper bound on the step loss.

or equivalently the constrained problem

$$\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right) \tag{4.5}$$

subject to[7]

$$\xi_i \geq 0 \qquad y_i \left(w^\top x_i + b\right) \geq 1 - \xi_i$$

(compare with (4.4)). See Figure 4.3.

Now let's write the Lagrangian for this problem, and solve it.

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i) \tag{4.6}$$

with dual variable constraints

$$\alpha_i \geq 0, \qquad \lambda_i \geq 0.$$

We minimize wrt the primal variables w, b, and ξ.

Derivative wrt w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \qquad w = \sum_{i=1}^{n}\alpha_i y_i x_i. \tag{4.7}$$

Derivative wrt b:

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0. \tag{4.8}$$

[7] To see this, we can write it as $\xi_i \geq 1 - y_i\left(w^\top x_i + b\right)$. Thus either $\xi_i = 0$, and $y_i\left(w^\top x_i + b\right) \geq 1$ as before, or $\xi_i > 0$, in which case to minimize (4.5), we'd use the smallest possible ξ_i satisfying the inequality, and we'd have $\xi_i = 1 - y_i\left(w^\top x_i + b\right)$.

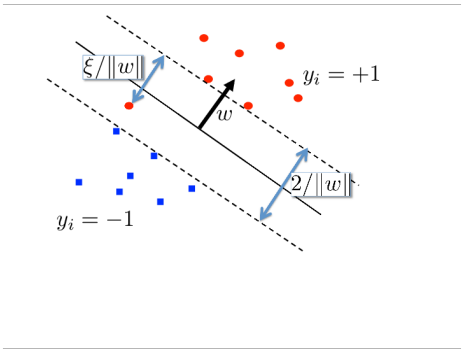

Figure 4.3: The nonseparable case. Note the red point which is a distance $\xi/\|w\|$ from the margin.

Derivative wrt ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \lambda_i = 0 \qquad \alpha_i = C - \lambda_i. \tag{4.9}$$

We can replace the final constraint by noting $\lambda_i \geq 0$, hence

$$\alpha_i \leq C.$$

Before writing the dual, we look at what these conditions imply about the scalars α_i that define the solution (4.7).

Non-margin SVs: $\alpha_i = C$:
Remember complementary slackness:

1. We immediately have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$.
2. Also, from condition $\alpha_i = C - \lambda_i$, we have $\lambda_i = 0$, hence it is possible that $\xi_i > 0$.

Margin SVs: $0 < \alpha_i < C$:

1. We again have $1 - \xi_i = y_i \left(w^\top x_i + b\right)$
2. This time, from $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

Non-SVs: $\alpha_i = 0$

1. This time we have: $y_i \left(w^\top x_i + b\right) > 1 - \xi_i$
2. From $\alpha_i = C - \lambda_i$, we have $\lambda_i \neq 0$, hence $\xi_i = 0$.

This means that the solution is sparse: all the points which are not either on the margin, or "margin errors", contribute nothing to the solution. In other words, only those points on the decision boundary, or which are margin errors,

contribute. Furthermore, the influence of the non-margin SVs is bounded, since their weight cannot exceed C: thus, severe outliers will not overwhelm the solution.

We now write the dual function, by substituting equations (4.7), (4.8), and (4.9) into (4.6), to get

$$\begin{aligned} g(\alpha,\lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left(w^\top x_i + b\right) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i) \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j - b\underbrace{\sum_{i=1}^{m}\alpha_i y_i}_{0} \\ &\qquad + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\underbrace{(C-\alpha_i)}_{\lambda_i}\xi_i \\ &= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j. \end{aligned}$$

Thus, our goal is to maximize the dual,

$$g(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \le \alpha_i \le C, \quad \sum_{i=1}^{n} y_i\alpha_i = 0.$$

So far we have defined the solution for w, but not for the offset b. This is simple to compute: for the margin SVs, we have $1 = y_i\left(w^\top x_i + b\right)$. Thus, we can obtain b from any of these, or take an average for greater numerical stability.

4.3 Kernelized version

We can straightforwardly define a maximum margin classifier in feature space. We write the original hinge loss formlation (ignoring the offset b for simplicity):

$$\min_w \left(\frac{1}{2}\|w\|_{\mathcal{H}}^2 + C\sum_{i=1}^{n}\left(1 - y_i\left\langle w, k(x_i,\cdot)\right\rangle_{\mathcal{H}}\right)_+\right)$$

for the RKHS $\mathcal{H}$ with kernel $k(x,\cdot)$. When we kernelize, we use the result of the representer theorem,

$$w = \sum_{i=1}^{n}\beta_i k(x_i,\cdot). \tag{4.10}$$

In this case, maximizing the margin is equivalent to minimizing $\|w\|_{\mathcal{H}}^2$: as we have seen, for many RKHSs (e.g. the RKHS corresponding to a Gaussian kernel), this corresponds to enforcing smoothness.

Substituting (4.10) and introducing the ξ_i variables, get

$$\min_{w,b} \left(\frac{1}{2}\beta^\top K\beta + C\sum_{i=1}^{n} \xi_i \right) \tag{4.11}$$

where the matrix K has i,jth entry $K_{ij} = k(x_i, x_j)$, subject to

$$\xi_i \geq 0 \qquad y_i \sum_{j=1}^{n} \beta_j k(x_i, x_j) \geq 1 - \xi_i$$

Thus, the primal variables w are replaced with β. The problem remains convex since K is positive definite. With some calculation (exercise!), the dual becomes

$$g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j k(x_i, x_j),$$

subject to the constraints

$$0 \leq \alpha_i \leq C,$$

and the decision function takes the form

$$w = \sum_{i=1}^{n} y_i \alpha_i k(x, \cdot).$$

4.4 The ν-SVM

It can be hard to interpret C. Therefore we modify the formulation to get a more intuitive parameter. Again, we drop b for simplicity. Solve

$$\min_{w,\rho,\xi} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\begin{aligned} \rho &\geq 0 \\ \xi_i &\geq 0 \\ y_i w^\top x_i &\geq \rho - \xi_i, \end{aligned}$$

where we see that we now optimize the margin width ρ. Thus, rather than choosing C, we now choose ν; the meaning of the latter will become clear shortly.

The Lagrangian is

$$\frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i - \nu\rho + \sum_{i=1}^{n} \alpha_i \left(\rho - y_i w^\top x_i - \xi_i\right) + \sum_{i=1}^{n} \beta_i(-\xi_i) + \gamma(-\rho)$$

for $\alpha_i \geq 0$, $\beta_i \geq 0$, and $\gamma \geq 0$. Differentiating wrt each of the primal variables w, ξ, ρ, and setting to zero, we get

$$\begin{aligned} w &= \sum_{i=1}^{n} \alpha_i y_i x_i \\ \alpha_i + \beta_i &= \frac{1}{n} \qquad (4.12) \\ \nu &= \sum_{i=1}^{n} \alpha_i - \gamma \qquad (4.13) \end{aligned}$$

From $\beta_i \geq 0$, equation (4.12) implies

$$0 \leq \alpha_i \leq n^{-1}.$$

From $\gamma \geq 0$ and (4.13), we get

$$\nu \leq \sum_{i=1}^{n} \alpha_i.$$

Let's now look at the complementary slackness conditions.
Assume $\rho > 0$ at the global solution, hence $\gamma = 0$, and (4.13) becomes

$$\sum_{i=1}^{n} \alpha_i = \nu. \qquad (4.14)$$

1. Case of $\xi_i > 0$: then complementary slackness states $\beta_i = 0$, hence from (4.12) we have $\alpha_i = n^{-1}$ for these points. Denote this set as $N(\alpha)$. Then

   $$\sum_{i \in N(\alpha)} \frac{1}{n} = \sum_{i \in N(\alpha)} \alpha_i \leq \sum_{i=1}^{n} \alpha_i = \nu,$$

   so

   $$\frac{|N(\alpha)|}{n} \leq \nu,$$

   and ν is an upper bound on the number of non-margin SVs.

2. Case of $\xi_i = 0$. Then $\alpha_i < n^{-1}$. Denote by $M(\alpha)$ the set of points $n^{-1} > \alpha_i > 0$. Then from (4.14),

   $$\nu = \sum_{i=1}^{n} \alpha_i = \sum_{i \in N(\alpha)} \frac{1}{n} + \sum_{i \in M(\alpha)} \alpha_i \leq \sum_{i \in M(\alpha) \cup N(\alpha)} \frac{1}{n},$$

   thus

   $$\nu \leq \frac{|N(\alpha)| + |M(\alpha)|}{n},$$

   and ν is a lower bound on the number of support vectors with non-zero weight (both on the margin, and "margin errors").

Substituting into the Lagrangian, we get

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \rho\nu - \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n}\alpha_i\rho - \sum_{i=1}^{n}\alpha_i\xi_i \\
&\qquad - \sum_{i=1}^{n}\left(\frac{1}{n} - \alpha_i\right)\xi_i - \rho\left(\sum_{i=1}^{n}\alpha_i - \nu\right) \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j
\end{aligned}$$

Thus, we must maximize

$$g(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j,$$

subject to

$$\sum_{i=1}^{n}\alpha_i \geq \nu \qquad 0 \leq \alpha_i \leq \frac{1}{n}.$$

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, England, 2004.

What is an RKHS?

Dino Sejdinovic, Arthur Gretton

March 11, 2012

1 Outline

- Normed and inner product spaces. Cauchy sequences and completeness. Banach and Hilbert spaces.
- Linearity, continuity and boundedness of operators. Riesz representation of functionals.
- Definition of an RKHS and reproducing kernels.
- Relationship with positive definite functions. Moore-Aronszajn theorem.

2 Some functional analysis

We start by reviewing some elementary Banach and Hilbert space theory. Two key results here will prove useful in studying the properties of reproducing kernel Hilbert spaces: (a) that a linear operator on a Banach space is continuous if and only if it is bounded, and (b) that all continuous linear functionals on a Banach space arise from the inner product. The latter is often termed *Riesz representation theorem*.

2.1 Definitions of Banach and Hilbert spaces

We will focus on *real* Banach and Hilbert spaces, which are, first of all, vector spaces[1] over the field $\mathbb{R}$ of real numbers. We remark that the theory remains valid in the context of *complex* Banach and Hilbert spaces, defined over the field $\mathbb{C}$ of complex numbers, modulo appropriately placed complex conjugates. In particular, the complex inner product satisfies conjugate symmetry instead of symmetry.

Definition 1 (Norm)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to [0, \infty)$ is said to be *a norm* on $\mathcal{F}$ if

1. $\|f\|_{\mathcal{F}} = 0$ if and only if $f = \mathbf{0}$ (*norm separates points*),

[1] A vector space can also be known as a linear space Kreyszig (1989, Definition 2.1-1).

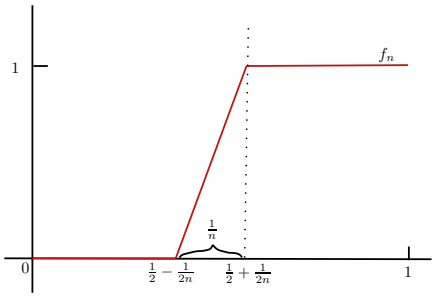

Figure 2.1: An example of a Cauchy sequence of continuous functions with no continuous limit, w.r.t. L_2-norm

2. $\|\lambda f\|_{\mathcal{F}} = |\lambda| \, \|f\|_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$ (*positive homogeneity*),

3. $\|f + g\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}} + \|g\|_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$ (*triangle inequality*).

Note that all elements in a normed space must have finite norm - if an element has infinite norm, it is not in the space. The norm $\|\cdot\|_{\mathcal{F}}$ induces a metric, i.e., a notion of distance on $\mathcal{F}$: $d(f, g) = \|f - g\|_{\mathcal{F}}$. This means that $\mathcal{F}$ is endowed with a certain topological structure, allowing us to study notions like continuity and convergence. In particular, we can consider when a sequence of elements of $\mathcal{F}$ converges with respect to induced distance. This gives rise to the definition of a convergent and of a Cauchy sequence:

Definition 2 (Convergent sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is said to *converge* to $f \in \mathcal{F}$ if for every $\epsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$, such that for all $n \geq N$, $\|f_n - f\|_{\mathcal{F}} < \epsilon$.

Definition 3 (Cauchy sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is said to be *a Cauchy (fundamental) sequence* if for every $\epsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$, such that for all $n, m \geq N$, $\|f_n - f_m\|_{\mathcal{F}} < \epsilon$.

From the triangle inequality $\|f_n - f_m\|_{\mathcal{F}} \leq \|f_n - f\|_{\mathcal{F}} + \|f - f_m\|_{\mathcal{F}}$, it is clear that every convergent sequence is Cauchy. However, not every Cauchy sequence in every normed space converges!

Example 4. The field of rational numbers $\mathbb{Q}$ with absolute value $|\cdot|$ as a norm is a normed vector space over itself. The sequence 1, 1.4, 1.41, 1.414, 1.4142, ... is a Cauchy sequence in $\mathbb{Q}$ which does not converge - because $\sqrt{2} \notin \mathbb{Q}$.

Example 5. In the space $C[0, 1]$ of bounded continuous functions on segment $[0, 1]$ endowed with the norm $\|f\| = \left(\int_0^1 |f(x)|^2 \, dx\right)^{1/2}$, a sequence $\{f_n\}$ of functions shown in Fig. 2.1, that take value 0 on $[0, \frac{1}{2} - \frac{1}{2n}]$ and value 1 on $[\frac{1}{2} + \frac{1}{2n}, 1]$ is Cauchy, but has no continuous limit.

Cauchy sequences are always bounded (Kreyszig, 1989, Exercise 4 p. 32), i.e., there exists $M < \infty$, s.t., $\|f_n\|_{\mathcal{F}} \leq M$, $\forall n \in \mathbb{N}$.

Next we define a complete space (Kreyszig, 1989, Definition 1.4-3):

Definition 6 (Complete space)**.** A space $\mathcal{X}$ is complete if every Cauchy sequence in $\mathcal{X}$ converges: it has a limit, and this limit is in $\mathcal{X}$.

Definition 7 (Banach space)**.** Banach space is a complete normed space, i.e., it contains the limits of all its Cauchy sequences.

In order to study useful geometrical notions analogous to those of Euclidean spaces $\mathbb{R}^d$, e.g., orthogonality, one requires additional structure on a Banach space, that is provided by a notion of inner product:

Definition 8 (Inner product)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be *an inner product* on $\mathcal{F}$ if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{F}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{F}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{F}}$
2. $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}$
3. $\langle f, f \rangle_{\mathcal{F}} \geq 0$ and $\langle f, f \rangle_{\mathcal{F}} = 0$ if and only if $f = 0$.

Vector space with an inner product is said to be an inner product (or unitary) space. Some immediate consequences of Definition 5 are that:

- $\langle 0, f \rangle_{\mathcal{F}} = 0, \ \forall f \in \mathcal{F}$,
- $\langle f, \alpha_1 g_1 + \alpha_2 g_2 \rangle_{\mathcal{F}} = \alpha_1 \langle f, g_1 \rangle_{\mathcal{F}} + \alpha_2 \langle f, g_2 \rangle_{\mathcal{F}}$.

One can always define *a norm* induced by the inner product:

$$\|f\|_{\mathcal{F}} = \langle f, f \rangle_{\mathcal{F}}^{1/2},$$

and the following useful relations between the norm and the inner product hold:

- $|\langle f, g \rangle| \leq \|f\| \cdot \|g\|$ (*Cauchy-Schwarz inequality*)
- $\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2$ (*the parallelogram law*)
- $4\langle f, g \rangle = \|f + g\|^2 - \|f - g\|^2$ (*the polarization identity*[2])

Definition 9 (Hilbert space)**.** Hilbert space is a complete inner product space, i.e., it is a Banach space with an inner product.

Two key examples of Hilbert spaces are given below.

[2] the polarization identity is different in complex Hilbert spaces and reads: $4\langle f, g \rangle = \|f + g\|^2 - \|f - g\|^2 + i\|f + ig\|^2 - i\|f - ig\|^2$

Example 10. For an index set A, the space $\ell^2(A)$ of sequences $\{x_\alpha\}_{\alpha\in A}$ of real numbers, satisfying $\sum_{\alpha\in A}|x_\alpha|^2<\infty$, endowed with the inner product

$$\langle\{x_\alpha\},\{y_\alpha\}\rangle_{\ell^2(A)}=\sum_{\alpha\in A}x_\alpha y_\alpha$$

is a Hilbert space.

Example 11. If μ is a positive measure on $\mathcal{X}\subset\mathbb{R}^d$, then the space

$$L_2(\mathcal{X};\mu):=\left\{f:X\to\mathbb{R}\text{ measurable}\;\middle|\;\|f\|_2=\left(\int_{\mathcal{X}}|f(x)|^2d\mu\right)^{1/2}<\infty\right\} \tag{2.1}$$

is a Hilbert space with inner product

$$\langle f,g\rangle=\int_X f(x)g(x)d\mu.$$

Strictly speaking, $L_2(\mathcal{X};\mu)$ is the space of equivalence classes of functions that differ by at most a set of μ-measure zero[3]. If μ is the Lebesgue measure, it is customary to write $L_2(\mathcal{X})$ as a shorthand[4].

More Hilbert space examples can be found in Kreyszig (1989, p. 132 and 133).

2.2 Bounded/Continuous linear operators

In the following, we take $\mathcal{F}$ and $\mathcal{G}$ to be normed vector spaces over $\mathbb{R}$ (for instance, they could both be the Banach spaces of functions mapping from $\mathcal{X}\subset\mathbb{R}$ to $\mathbb{R}$, with L_p-norm)

Definition 12 (Linear operator)**.** A function $A:\mathcal{F}\to\mathcal{G}$, where $\mathcal{F}$ and $\mathcal{G}$ are both normed linear spaces over $\mathbb{R}$, is called a **linear** operator if and only if it satisfies the following properties:

- **Homogeneity:** $A(\alpha f)=\alpha(Af)\quad\forall\alpha\in\mathbb{R},\,f\in\mathcal{F}$,
- **Additivity:** $A(f+g)=Af+Ag\quad\forall f,g\in\mathcal{F}$.

Example 13. Let $\mathcal{F}$ be an inner product space. For $g\in\mathcal{F}$, operator $A_g:\mathcal{F}\to\mathbb{R}$, defined with $A_g(f):=\langle f,g\rangle_{\mathcal{F}}$ is a linear operator. Note that the image space of A_g is the underlying field $\mathbb{R}$, which is trivially a normed linear space over itself[5]. Such scalar-valued operators are called *functionals* on $\mathcal{F}$.

[3] Norm defined in (2.1) does not separate functions f and g which differ only on some set $A\subset\mathcal{X}$, for which $\mu(A)=0$, since $f-g\neq 0$ and $\|f-g\|_2^2=\int_{\mathcal{X}}\left(|f(x)-g(x)|^2\right)d\mu=0$. Thus, we consider all such functions as a single element in the space $L_2(\mathcal{X};\mu)$.

[4] In fact, $\ell^2(A)$ is just $L_2(\mathcal{X};\mu)$ where μ is the *counting measure*, where the "size" of a subset is taken to be the number of elements in the subset

[5] with norm $|\cdot|$

Definition 14 (Continuity)**.** A function $A : \mathcal{F} \to \mathcal{G}$ is said to be **continuous** at $f_0 \in \mathcal{F}$, if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon, f_0) > 0$, s.t.

$$\|f - f_0\|_{\mathcal{F}} < \delta \qquad \text{implies} \qquad \|Af - Af_0\|_{\mathcal{G}} < \epsilon. \tag{2.2}$$

A is **continuous** on $\mathcal{F}$, if it is continuous at every point of $\mathcal{F}$. If, in addition, δ depends on ϵ only, i.e., $\forall \epsilon > 0$, $\exists \delta = \delta(\epsilon) > 0$, s.t. (2.2) holds for every $f_0 \in \mathcal{F}$, A is said to be **uniformly continuous**.

In other words, continuity means that a convergent sequence in $\mathcal{F}$ is mapped to a convergent sequence in $\mathcal{G}$. An even stronger form of continuity than uniform continuity is Lipschitz continuity:

Definition 15 (Lipschitz continuity)**.** A function $A : \mathcal{F} \to \mathcal{G}$ is said to be **Lipschitz continuous** if $\exists C > 0$, s.t. $\forall f_1, f_2 \in \mathcal{F}$, $\|Af_1 - Af_2\|_{\mathcal{G}} \leq C \|f_1 - f_2\|_{\mathcal{F}}$.

It is clear that Lipschitz continuous function is uniformly continuous since one can choose $\delta = \epsilon / C$.

Example 16. For $g \in \mathcal{F}$, $A_g : \mathcal{F} \to \mathbb{R}$, defined with $A_g(f) := \langle f, g \rangle_{\mathcal{F}}$ is continuous on $\mathcal{F}$:

$$|A_g(f_1) - A_g(f_2)| = |\langle f_1 - f_2, g \rangle_{\mathcal{F}}| \leq \|g\|_{\mathcal{F}} \|f_1 - f_2\|_{\mathcal{F}} .$$

Definition 17 (Operator norm)**.** The operator norm of a linear operator $A : \mathcal{F} \to \mathcal{G}$ is defined as

$$\|A\| = \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$$

Definition 18 (Bounded operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ is said to be a bounded operator if $\|A\| < \infty$.

It can readily be shown (Kreyszig, 1989) that operator norm satisfies all the requirements of a norm (triangle inequality, zero iff the operator maps only to the zero function, $\|\lambda A\| = |\lambda| \|A\|$ for $c \in \mathbb{R}$), and that the set of bounded linear operators $A : \mathcal{F} \to \mathcal{G}$ (for which the norm is defined) is therefore itself a normed vector space. Another way to write the above is to say that, for $f \in \mathcal{F}$ (possibly) not attaining the supremum, we have

$$\begin{aligned} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}} &\leq \|A\| \\ \|Af\|_{\mathcal{G}} &\leq \|A\| \|f\|_{\mathcal{F}} , \end{aligned}$$

so there exists a non-negative real number λ for which $\|Af\|_{\mathcal{G}} \leq \lambda \|f\|_{\mathcal{F}}$, for all $f \in \mathcal{F}$, and the **smallest** such λ is precisely the operator norm. In other words, a bounded subset in $\mathcal{F}$ is mapped to a bounded subset in $\mathcal{G}$.

WARNING: In calculus, a bounded function is a function whose range is a bounded set. This definition is *not* the same as the above, which simply states that the effect of A on f is bounded by some scaling of the norm of f. There is

a useful geometric interpretation of the operator norm: A maps the closed unit ball in $\mathcal{F}$, into a subset of the closed ball in $\mathcal{G}$ centered at $0 \in \mathcal{G}$ and with radius $\|A\|$. Note also the result in Kreyszig (1989, p. 96): every linear operator on a normed, finite dimensional space is bounded.

Theorem 19. *Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ be normed linear spaces. If L* is a linear operator, then the following three conditions are equivalent:

1. *L is a bounded operator.*
2. *L is continuous on $\mathcal{F}$.*
3. *L is continuous at one point of $\mathcal{F}$.*

Proof. (1)⇒(2), since $\|L(f_1 - f_2)\|_{\mathcal{G}} \leq \|L\| \, \|f_1 - f_2\|_{\mathcal{F}}$, L is Lipschitz continuous with a Lipschitz constant $\|L\|$, and (2)⇒(3) trivially. Now assume that L is continuous at one point $f_0 \in \mathcal{F}$. Then, there is a $\delta > 0$, s.t. $\|L\Delta\|_{\mathcal{G}} = \|L(f_0 + \Delta) - Lf_0)\|_{\mathcal{G}} \leq 1$, whenever $\|\Delta\|_{\mathcal{F}} \leq \delta$. But then, $\forall f \in \mathcal{F} \backslash \{0\}$, since $\left\| \delta \frac{f}{\|f\|} \right\|_{\mathcal{F}} = \delta$,

$$\begin{aligned} \|Lf\|_{\mathcal{G}} &= \delta^{-1} \|f\|_{\mathcal{F}} \left\| L\left(\delta \frac{f}{\|f\|} \right) \right\|_{\mathcal{G}} \\ &\leq \delta^{-1} \|f\|_{\mathcal{F}}, \end{aligned}$$

so $\|L\| \leq \delta^{-1}$, and (3)⇒(1), q.e.d. □

Definition 20 (Algebraic dual)**.** If $\mathcal{F}$ is a normed space, then the space $\mathcal{F}'$ of *linear* functionals $A : \mathcal{F} \to \mathbb{R}$ is called the algebraic dual space of $\mathcal{F}$.

Definition 21 (Topological dual)**.** If $\mathcal{F}$ is a normed space, then the space $\mathcal{F}'$ of *continuous linear* functionals $A : \mathcal{F} \to \mathbb{R}$ is called the topological dual space of $\mathcal{F}$.

In finite-dimensional space, the two notions of dual spaces coincide. However, this is not the case in infinite dimensions. Unless otherwise specified, we refer to the topological dual when discussing the dual of $\mathcal{F}$.

We have seen in Examples 13, 16 that the functionals of the form $\langle \cdot, g \rangle_{\mathcal{F}}$ on an inner product space $\mathcal{F}$ are both linear and continuous, i.e., they lie in the topological dual $\mathcal{F}'$ of $\mathcal{F}$. It turns out that if $\mathcal{F}$ is a Hilbert space, all elements of $\mathcal{F}'$ take this form[6].

Theorem 22. (Riesz representation) *In a Hilbert space $\mathcal{F}$, all continous linear functionals are of the form $\langle \cdot, g \rangle_{\mathcal{F}}$, for some $g \in \mathcal{F}$.*

Two Hilbert spaces may have elements of different nature, e.g., functions vs. sequences, but still have exactly the same geometric structure. This is the notion of *isometric isomorphism* of two Hilbert spaces. It combines notions of

[6] An approachable proof of Riesz representation theorem is in Rudin (1987, Theorem 4.12).

vector space isomorphism (a linear bijection) and of *isometry* (transformation that preserves distances). Once the isometric isomorphism of two spaces is established, it is customary to use whichever of the spaces is more convenient for the problem.

Definition 23 (Hilbert space isomorphism)**.** Two Hilbert spaces $\mathcal{H}$ and $\mathcal{F}$ are said to be *isometrically isomorphic* if there is a linear bijective map $U : \mathcal{H} \to \mathcal{F}$, which preserves the inner product, i.e., $\langle h_1, h_2 \rangle_{\mathcal{H}} = \langle Uh_1, Uh_2 \rangle_{\mathcal{F}}$.

Note that Riesz representation theorem gives us a natural isometric isomorphism[7] $\psi : g \mapsto \langle \cdot, g \rangle_{\mathcal{F}}$ between $\mathcal{F}$ and $\mathcal{F}'$, whereby $\|\psi(g)\|_{\mathcal{F}'} = \|g\|_{\mathcal{F}}$. This property will be used below when defining a kernel on RKHSs. In particular, note that the dual space of a Hilbert space is also a Hilbert space.

3 Reproducing kernel Hilbert space

3.1 Definition of an RKHS

We begin by describing in general terms the reproducing kernel Hilbert space, and its associated kernel. Let $\mathcal{H}$ be a Hilbert space[8] of functions mapping from some non-empty set $\mathcal{X}$ to $\mathbb{R}$ (we write this: $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$). A very interesting property of an RKHS is that if two functions $f \in \mathcal{H}$ and $g \in \mathcal{H}$ are close in the norm of $\mathcal{H}$, then $f(x)$ and $g(x)$ are close for all $x \in \mathcal{X}$. We write the inner product on $\mathcal{H}$ as $\langle f, g \rangle_{\mathcal{H}}$, and the associated norm $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$. We may alternatively write the function f as $f(\cdot)$, to indicate it takes an argument in $\mathcal{X}$.

Note that since $\mathcal{H}$ is now a space of functions on $\mathcal{X}$, there is for every $x \in \mathcal{X}$ a very special functional on $\mathcal{H}$: the one that assigns to each $f \in \mathcal{H}$, its value at x:

Definition 24 (Evaluation functional)**.** Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x : f \mapsto f(x)$ is called the (Dirac) evaluation functional at x.

It is clear that evaluation functionals are always linear: For $f, g \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$, $\delta_x(\alpha f + \beta g) = (\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x) = \alpha \delta_x(f) + \beta \delta_x(g)$. So the natural question is whether they are also continuous (recall that this is the same as bounded). This is exactly how reproducing kernel Hilbert space are defined (Steinwart & Christmann, 2008, Definition 4.18(ii)):

Definition 25 (Reproducing kernel Hilbert space)**.** A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if δ_x is continuous $\forall x \in \mathcal{X}$.

A useful consequence is that RKHSs are particularly well behaved, relative to other Hilbert spaces.

[7] in complex Hilbert spaces, due to conjugate symmetry of inner product, this map is antilinear, i.e., $\psi(\alpha g) = \bar{\alpha}\,(\psi g)$

[8] This is a complete linear space with a dot product - see earlier.

Corollary 26. *(Norm convergence in $\mathcal{H}$ implies pointwise convergence)(Berlinet & Thomas-Agnan, 2004, Corollary 1) If two functions converge in RKHS norm, then they converge at every point, i.e., if* $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, *then* $\lim_{n\to\infty} f_n(x) = f(x)$, $\forall x \in \mathcal{X}$.

Proof. For any $x \in \mathcal{X}$,

$$\begin{aligned} |f_n(x) - f(x)| &= |\delta_x f_n - \delta_x f| \\ &\leq \|\delta_x\| \, \|f_n - f\|_{\mathcal{H}} , \end{aligned}$$

where $\|\delta_x\|$ is the norm of the evaluation operator (which is bounded by definition on the RHKS). □

Example 27. (Berlinet & Thomas-Agnan, 2004, p. 2) If we are *not* in an RKHS, then norm convergence does not necessarily imply pointwise convergence. Let $\mathcal{H} = L_2([0,1])$, endowed with the metric

$$\|f_1 - f_2\|_{L_2([0,1])} = \left(\int_0^1 |f_1(x) - f_2(x)|^2 \, dx \right)^{1/2},$$

and consider the sequence of functions $\{q_n\}_{n=1}^{\infty}$, where $q_n = x^n$. Then

$$\begin{aligned} \lim_{n\to\infty} \|q_n - 0\|_{L_2([0,1])} &= \lim_{n\to\infty} \left(\int_0^1 x^{2n} dx \right)^{1/2} \\ &= \lim_{n\to\infty} \frac{1}{\sqrt{2n+1}} \\ &= 0, \end{aligned}$$

and yet $q_n(1) = 1$ for all n, i.e., $q_n \to 0 \in \mathcal{H}$, but[9] $q_n(1) \nrightarrow 0$. In other words, the evaluation of functions at point 1 is not continuous on the set $\{q_n\}_{n=1}^{\infty}$.

3.2 Reproducing kernels

The reader will note that there is no mention of a kernel in the definition of an RKHS! We next define what is meant by a kernel, and then show how it fits in with the above definition.

Definition 28. (Reproducing kernel (Berlinet & Thomas-Agnan, 2004, p. 7))

Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a reproducing kernel* of $\mathcal{H}$ if it satisfies

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

[9] A careful reader will note that this arises precisely because $L_2([0,1])$ is a space in which a function that is zero everywhere and a function that is zero everywhere apart from a finite number of points are treated as the same element!

In particular, for any $x, y \in \mathcal{X}$,

$$k(x,y) = \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}. \tag{3.1}$$

The definition above raises a number of questions. What does the kernel have to do with the definition of the RKHS? Does this kernel exist? Is it unique? What properties does it have? We first consider uniqueness, which is immediate from the definition of the reproducing kernel.

Proposition 29. (Uniqueness of the reproducing kernel) *If it exists, reproducing kernel is unique.*

Proof. Assume that $\mathcal{H}$ has two reproducing kernels k_1 and k_2. Then,

$$\langle f, k_1(\cdot,x) - k_2(\cdot,x)\rangle_{\mathcal{H}} = f(x) - f(x) = 0,\ \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

In particular, if we take $f = k_1(\cdot,x) - k_2(\cdot,x)$, we obtain $\|k_1(\cdot,x) - k_2(\cdot,x)\|_{\mathcal{H}}^2 = 0$, $\forall x \in \mathcal{X}$, i.e., $k_1 = k_2$. □

To establish existence of a reproducing kernel in an RKHS, we will make use of the Riesz representation theorem - which tells us that in an RKHS, evaluation itself can be represented as an inner product!

Theorem 30. (Existence of the reproducing kernel) *$\mathcal{H}$ is a reproducing kernel Hilbert space (i.e., its evaluation functionals δ_x are continuous linear operators), if and only if $\mathcal{H}$ has a reproducing kernel.*

Proof. Given that a Hilbert space $\mathcal{H}$ has a reproducing kernel k with the reproducing property $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = f(x)$, then

$$\begin{aligned} |\delta_x f| &= |f(x)| \\ &= |\langle f, k(\cdot,x)\rangle_{\mathcal{H}}| \\ &\leq \|k(\cdot,x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \langle k(\cdot,x), k(\cdot,x)\rangle_{\mathcal{H}}^{1/2} \|f\|_{\mathcal{H}} \\ &= k(x,x)^{1/2} \|f\|_{\mathcal{H}} \end{aligned}$$

where the third line uses the Cauchy-Schwarz inequality. Consequently, $\delta_x : \mathcal{F} \to \mathbb{R}$ is a bounded linear operator.

To prove the other direction, assume that $\delta_x \in \mathcal{H}'$, i.e. $\delta_x : \mathcal{F} \to \mathbb{R}$ is a bounded linear functional. The Riesz representation theorem (Theorem 22) states that there exists an element $f_{\delta_x} \in \mathcal{H}$ such that

$$\delta_x f = \langle f, f_{\delta_x}\rangle_{\mathcal{H}},\ \forall f \in \mathcal{H}.$$

Define $k(x',x) = f_{\delta_x}(x')$, $\forall x, x' \in \mathcal{X}$. Then, clearly $k(\cdot,x) = f_{\delta_x} \in \mathcal{H}$, and $\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = \delta_x f = f(x)$. Thus, k is the reproducing kernel. □

From the above, we see $k(\cdot, x)$ is in fact the *representer of evaluation* at x. We now turn to one of the most important properties of the kernel function: specifically, that it is positive definite (Berlinet & Thomas-Agnan, 2004, Definition 2), (Steinwart & Christmann, 2008, Definition 4.15).

Definition 31 (Positive definite functions)**.** A symmetric[10] function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, \dots a_n) \in \mathbb{R}^n$, $\forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0. \tag{3.2}$$

The function $h(\cdot, \cdot)$ is *strictly* positive definite if for mutually distinct x_i, the equality holds only when all the a_i are zero.[11]

Every inner product is a positive definite function, and more generally:

Lemma 32. *Let $\mathcal{F}$ be any Hilbert space (not necessarily an RKHS), $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{F}$. Then $h(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$ is a positive definite function.*

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{F}}^2 \geq 0.
\end{aligned}$$

□

Corollary 33. *Reproducing kernels are positive definite.*

Proof. For a reproducing kernel k in an RKHS $\mathcal{H}$, one has $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$, so it is sufficient to take $\phi : x \mapsto k(\cdot, x)$. □

[10] Note that we require symmetry of h in addition to (3.2). In the complex case, $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \bar{a_j} h(x_i, x_j) \geq 0$, satisfied for all complex scalars $(a_1, \dots a_n) \in \mathbb{C}^n$ will itself imply conjugate symmetry of h. Can you construct a non-symmetric h that satisfies (3.2)?

[11] Note that Wendland (2005, Definition 6.1 p. 65) uses the terminology "positive semi-definite" vs "positive definite". This is probably more logical, since it then coincides with the terminology used in linear algebra. However, we proceed with the terminology prevalent with machine learning literature.

3.3 Feature space, and other kernel properties

This section summarizes the relevant parts of Steinwart & Christmann (2008, Section 4.1).

Following Lemma 32, one can define a *kernel* (note that we drop qualification *reproducing* here - later we will see that these two notions are the same), as a function which can be represented via inner product, and this is the approach taken in Steinwart & Christmann (2008, Section 4.1):

Definition 34 (*Kernel*)**.** Let $\mathcal{X}$ be a non-empty set. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *kernel* if there exists a real Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{H}$,

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} .$$

Such map $\phi : \mathcal{X} \to \mathcal{H}$ is referred to as the feature map, and space $\mathcal{H}$ as the feature space. For a given kernel, there may be more than one feature map, as demonstrated by the following example.

Example 35. Consider $\mathcal{X} = \mathbb{R}$, and

$$k(x, y) = xy = \begin{bmatrix} \frac{x}{\sqrt{2}} & \frac{x}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{y}{\sqrt{2}} \\ \frac{y}{\sqrt{2}} \end{bmatrix},$$

where we defined the feature maps $\phi(x) = x$ and $\tilde{\phi}(x) = \begin{bmatrix} \frac{x}{\sqrt{2}} & \frac{x}{\sqrt{2}} \end{bmatrix}$, and where the feature spaces are respectively, $\mathcal{H} = \mathbb{R}$, and $\tilde{\mathcal{H}} = \mathbb{R}^2$.

Lemma 36. [ℓ_2 convergent sequences are kernel feature maps] *For every $x \in \mathcal{X}$, assume the sequence $\{f_n(x)\} \in \ell_2$ for $n \in \mathbb{N}$, where $f_n : \mathcal{X} \to \mathbb{R}$. Then*

$$k(x_1, x_2) := \sum_{n=1}^{\infty} f_n(x_1) f_n(x_2) \tag{3.3}$$

is a kernel.

Proof. Hölder's inequality states

$$\sum_{n=1}^{\infty} |f_n(x_1) f_n(x_2)| \leq \|f_n(x_1)\|_{\ell_2} \|f_n(x_2)\|_{\ell_2} .$$

so the series (3.3) converges absolutely. Definining $\mathcal{H} := \ell_2$ and $\phi(x) = \{f_n(x)\}$ completes the proof. □

4 Construction of an RKHS from a kernel: Moore-Aronsajn

We have seen previously that *given* a reproducing kernel Hilbert space $\mathcal{H}$, we may define a unique reproducing kernel associated with $\mathcal{H}$, which is a positive

definite function. Then we considered kernels, i.e., functions that can be written as an inner product in a feature space. All reproducing kernels are kernels. In Example (35), we have seen that the representation of a kernel as an inner product in a feature space may not be unique. However, neither of the feature spaces in that example is an RKHS, as they are not spaces of functions on $\mathcal{X} = \mathbb{R}$.

Our goal now is to show that for every positive definite function $k(x, y)$, there corresponds a *unique RKHS* $\mathcal{H}$, for which k is a reproducing kernel. The proof is rather tricky, but also very revealing of the properties of RKHSs, so it is worth understanding (it also occurs in very incomplete form in a number of books and tutorials, so it is worth seeing what a complete proof looks like).

Starting with the kernel, we will construct a pre-RKHS $\mathcal{H}_0$, from which we will form the RKHS $\mathcal{H}$. The pre-RKHS $\mathcal{H}_0$ must satisfy two properties:

1. the evaluation functionals δ_x are continuous on $\mathcal{H}_0$,
2. Any Cauchy sequence f_n in $\mathcal{H}_0$ which converges pointwise to 0 also converges in $\mathcal{H}_0$-norm to 0.

The last result has an important implication: Any Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0$ that converges pointwise to $f \in \mathcal{H}_0$, also converges to f in $\|\cdot\|_{\mathcal{H}_0}$, since in that case $\{f_n - f\}$ converges pointwise to 0, and thus $\|f_n - f\|_{\mathcal{H}_0} \to 0$. **PREVIEW**: we can already say what the pre-RKHS $\mathcal{H}_0$ will look like: it is the set of functions

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x). \tag{4.1}$$

After the proof, we'll show in Section (4.5) that these functions satisfy conditions (1) and (2) of the pre-Hilbert space.

Next, **define** $\mathcal{H}$ to be the set of functions $f \in \mathbb{R}^{\mathcal{X}}$ for which there exists an $\mathcal{H}_0$-Cauchy sequence $\{f_n\} \in \mathcal{H}_0$ converging pointwise to f: note that $\mathcal{H}_0 \subset \mathcal{H}$, since the limits of these Cauchy sequences might not be in $\mathcal{H}_0$. Our goal is to prove that $\mathcal{H}$ is an RKHS. The two properties above hold if and only if

- $\mathcal{H}_0 \subset \mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ and the topology induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ on $\mathcal{H}_0$ coincides with the topology induced on $\mathcal{H}_0$ by $\mathcal{H}$.
- $\mathcal{H}$ has reproducing kernel $k(x, y)$.

We concern ourselves with proving that (1), (2) imply the above bullet points, since the reverse direction is easy to prove. This takes four steps:

1. We define the inner product between $f, g \in \mathcal{H}$ as the limit of an inner product of the Cauchy sequences $\{f_n\}$, $\{g_n\}$ converging to f and g respectively. Is the inner product well defined, and independent of the sequences used? This is proved in Section 4.1.
2. Recall that an inner product space must satisfy $\langle f, f \rangle_{\mathcal{H}} = 0$ iff $f = 0$. Is this true when we define the inner product on $\mathcal{H}$ as above? (Note that we

can also check that the remaining requirements for an inner product on $\mathcal{H}$ hold, but these are straightforward)

3. Are the evaluation functionals still continuous on $\mathcal{H}$?

4. Is $\mathcal{H}$ complete? I.e., is it a Hilbert space?

Finally, we'll see that the functions (4.1) define a valid pre-RKHS $\mathcal{H}_0$. We will also show that the kernel $k(\cdot, x)$ has the reproducing property on the RKHS $\mathcal{H}$.

4.1 Is the inner product well defined in $\mathcal{H}$?

In this section we prove that if we define the inner product in $\mathcal{H}$ of all limits of Cauchy sequences as (4.2) below, then this limit is *well defined*: (1) it converges, and (2) it depends only on the *limits* of the Cauchy sequences, and not the particular sequences themselves.

This result is from Berlinet & Thomas-Agnan (2004, Lemma 5).

Lemma 37. *For $f, g \in \mathcal{H}$ and Cauchy sequences (wrt the $\mathcal{H}_0$ norm) $\{f_n\}$, $\{g_n\}$ converging pointwise to f and g, define $\alpha_n = \langle f_n, g_n \rangle_{\mathcal{H}_0}$. Then, $\{\alpha_n\}$ is convergent and its limit depends only on f and g. We thus define*

$$\langle f, g \rangle_{\mathcal{H}} := \lim_{n \to \infty} \langle f_n, g_n \rangle_{\mathcal{H}_0} \tag{4.2}$$

Proof that $\alpha_n = \langle f_n, g_n \rangle_{\mathcal{H}_0}$ is convergent: For $n, m \in \mathbb{N}$,

$$\begin{aligned}
|\alpha_n - \alpha_m| &= \left|\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{H_0}\right| \\
&\quad \left|\langle f_n, g_n \rangle_{H_0} - \langle f_m, g_n \rangle_{H_0} + \langle f_m, g_n \rangle_{H_0} - \langle f_m, g_m \rangle_{H_0}\right| \\
&= \left|\langle f_n - f_m, g_n \rangle_{H_0} + \langle f_m, g_n - g_m \rangle_{H_0}\right| \\
&\leq \left|\langle f_n - f_m, g_n \rangle_{H_0}\right| + \left|\langle f_m, g_n - g_m \rangle_{H_0}\right| \\
&\leq \|g_n\|_{H_0} \|f_n - f_m\|_{H_0} + \|f_m\|_{H_0} \|g_n - g_m\|_{H_0} .
\end{aligned}$$

Take $\epsilon > 0$. Every Cauchy sequence is bounded, so $\exists A, B \in \mathbb{R}$, $\|f_m\|_{H_0} \leq A$, $\|g_n\|_{H_0} \leq B$, $\forall n, m \in \mathbb{N}$.

By taking $N_1 \in \mathbb{N}$ s.t. $\|f_n - f_m\|_{H_0} < \frac{\epsilon}{2B}$, for $n, m \geq N_1$, and $N_2 \in \mathbb{N}$ s.t. $\|g_n - g_m\|_{H_0} < \frac{\epsilon}{2A}$, for $n, m \geq N_1$, we have that $|\alpha_n - \alpha_m| < \epsilon$, for $n, m \geq \max(N_1, N_2)$, which means that $\{\alpha_n\}$ is a Cauchy sequence in $\mathbb{R}$, which is complete, and the sequence is convergent in $\mathbb{R}$.

Proof that limit is independent of Cauchy sequence chosen:

If some $\mathcal{H}_0$-Cauchy sequences $\{f'_n\}$, $\{g'_n\}$ also converge pointwise to f and g, and $\alpha'_n = \langle f'_n, g'_n \rangle_{H_0}$, one similarly shows that

$$|\alpha_n - \alpha'_n| \leq \|g_n\|_{H_0} \|f_n - f'_n\|_{H_0} + \|f'_n\|_{H_0} \|g_n - g'_n\|_{H_0} .$$

Now, since $\{f_n\}$ and $\{f'_n\}$ both converge pointwise to f, $\{f_n - f'_n\}$ converges pointwise to 0, and so does $\{g_n - g'_n\}$. But then they also converge to 0 in $\|\cdot\|_{H_0}$ by the pre-RKHS axiom 2, and therefore $\{\alpha_n\}$ and $\{\alpha'_n\}$ must have the same limit.

4.2 Does it hold that $\langle f, f \rangle_{\mathcal{H}} = 0$ iff $f = 0$?

In this section, we verify that all the expected properties of an inner product from Definition (8) hold for $\mathcal{H}$. It turns out that the only challenging property to show is the third one - the others follow from the inner product definition on the pre-RKHS. This is **?**, Lemma 6.

Lemma 38. *Let $\{f_n\}$ be Cauchy sequence in $\mathcal{H}_0$ converging pointwise to $f \in \mathcal{H}$. If $\lim_{n\to\infty} \langle f_n, f_n \rangle_{\mathcal{H}_0} = \|f_n\|^2_{\mathcal{H}_0} = 0$, then $f(x) = 0$ pointwise for all x (we assumed pointwise convergence implies norm convergence - we now want to prove the other direction, bearing in mind that the inner product in $\mathcal{H}$ is defined as the limit of inner products in $\mathcal{H}_0$ by (4.2)).*

Proof: $\forall x \in \mathcal{X}$, $f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) \underset{\text{(a)}}{\leq} \lim_{n\to\infty} \|\delta_x\| \, \|f_n\|_{\mathcal{H}_0} \underset{\text{(b)}}{=} 0$, where in (a) we used that the evaluation functional δ_x is continuous on $\mathcal{H}_0$, by the pre-RKHS axiom 1 (hence bounded, with a well defined operator norm $\|\delta_x\|$); and in (b) we used the assumption in the lemma that f_n converges to 0 in $\|\cdot\|_{\mathcal{H}_0}$.

4.3 Are the evaluation functionals continuous on $\mathcal{H}$?

Here we need to establish a preliminary lemma, before we can continue.

Lemma 39. *$\mathcal{H}_0$ is dense in $\mathcal{H}$ (Berlinet & Thomas-Agnan, 2004, Lemma 7, Corollary 2).*

Proof. It suffices to show that given any $f \in \mathcal{H}$ and its associated Cauchy sequence $\{f_n\}$ wrt $\mathcal{H}_0$ converging pointwise to f (which exists by definition), $\{f_n\}$ also converges to f in $\|\cdot\|_{\mathcal{H}}$ (note: this is the *new* norm which we defined above in terms of limits of Cauchy sequences in $\mathcal{H}_0$).

Since$\{f_n\}$ is Cauchy in $\mathcal{H}_0$-norm, for all $\epsilon > 0$, there is $N \in \mathbb{N}$, s.t. $\|f_m - f_n\|_{\mathcal{H}_0} < \epsilon$, $\forall m, n \geq N$. Fix $n^* \geq N$. The sequence $\{f_m - f_{n^*}\}_{m=1}^{\infty}$ converges pointwise to $f - f_{n^*}$. We now simply use the definition of the inner product in $\mathcal{H}$ from (4.2),

$$\|f - f_{n^*}\|^2_{\mathcal{H}} \quad = \quad \lim_{m\to\infty} \|f_m - f_{n^*}\|^2_{\mathcal{H}_0} \leq \epsilon^2,$$

whereby $\{f_n\}_{n=1}^{\infty}$ converges to f in $\|\cdot\|_{\mathcal{H}}$. □

Lemma 40. *The evaluation functionals are continuous on $\mathcal{H}$ (Berlinet & Thomas-Agnan, 2004, Lemma 8).*

Proof. We show that δ_x is continuous at $f = 0$, since this implies by linearity that it is continuous everywhere. Let $x \in \mathcal{X}$, and $\epsilon > 0$. By pre-RKHS axiom 1, δ_x is continuous on $\mathcal{H}_0$. Thus, $\exists \eta$, s.t.

$$\|g - 0\|_{\mathcal{H}_0} = \|g\|_{\mathcal{H}_0} < \eta \Rightarrow |\delta_x(g)| = |g(x)| < \epsilon/2. \tag{4.3}$$

To complete the proof, we just need to show that there is a $g \in \mathcal{H}_0$ close (in $\mathcal{H}$-norm) to some $f \in \mathcal{H}$ with small norm, and that this function is also close at each point.

We take $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} < \eta/2$. By Lemma (39) there is a Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0$ converging both pointwise to f and in $\|\cdot\|_{\mathcal{H}}$ to f, so one can find $N \in \mathbb{N}$, s.t.

$$\begin{aligned} |f(x) - f_N(x)| &< \epsilon/2, \\ \|f - f_N\|_{\mathcal{H}} &< \eta/2. \end{aligned}$$

We have from these definitions that

$$\|f_N\|_{\mathcal{H}_0} = \|f_N\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} + \|f - f_N\|_{\mathcal{H}} < \eta.$$

Thus $\|f\|_{\mathcal{H}} < \eta/2$ implies $\|f_N\|_{\mathcal{H}_0} < \eta$. Using (4.3) and setting $g := f_N$, we have that $\|f_N\|_{\mathcal{H}_0} < \eta$ *implies* $|f_N(x)| < \epsilon/2$, and thus $|f(x)| \leq |f(x) - f_N(x)| + |f_N(x)| < \epsilon$. In other words, $\|f\|_{\mathcal{H}} < \eta/2$ is shown to imply $|f(x)| < \epsilon$. This means that δ_x is continuous at 0 in the $\|\cdot\|_{\mathcal{H}}$ sense, and thus by linearity on all $\mathcal{H}$. □

4.4 Is $\mathcal{H}$ complete (a Hilbert space)?

The idea here is to show that every Cauchy sequence wrt the $\mathcal{H}$-norm converges to a function in $\mathcal{H}$.

Lemma 41. *$\mathcal{H}$ is complete.*

Let $\{f_n\}$ be any Cauchy sequence in $\mathcal{H}$. Since evaluation functionals are linear continuous on $\mathcal{H}$ by (40), then for any $t \in E$, $\{f_n(t)\}$ is convergent in $\mathbb{R}$ to some $f(t) \in \mathbb{R}$ (since $\mathbb{R}$ is complete, it contains this limit). The question is thus whether the function $f(t)$ defined pointwise in this way is still in $\mathcal{H}$ (recall that $\mathcal{H}$ is defined as containing the limit of $\mathcal{H}_0$-Cauchy sequences that converge pointwise).

The proof strategy is to define a sequence of functions$\{g_n\}$, where $g_n \in \mathcal{H}_0$, which is "close" to the $\mathcal{H}$-Cauchy sequence $\{f_n\}$. These functions will then be shown **(1)** to converge pointwise to f, and **(2)** to be Cauchy in $\mathcal{H}_0$. Hence by our original construction of $\mathcal{H}$, we have $f \in \mathcal{H}$. Finally, we show $f_n \to f$ in $\mathcal{H}$-norm.

Define $f(x) := \lim_{n\to\infty} f_n(x)$. For $n \in \mathbb{N}$, choose $g_n \in \mathcal{H}_0$ such that $\|g_n - f_n\|_{\mathcal{H}} < \frac{1}{n}$. This can be done since $\mathcal{H}_0$ is dense in $\mathcal{H}$. From

$$\begin{aligned} |g_n(x) - f(x)| &\leq |g_n(x) - f_n(x)| + |f_n(x) - f(x)| \\ &\leq |\delta_x(g_n - f_n)| + |f_n(x) - f(x)|, \end{aligned}$$

The first term in this sum goes to zero due to the continuity of δ_x on $\mathcal{H}$ (Lemma (40)), and thus$\{g_n(x)\}$ converges to $f(x)$, satisfying criterion (1). For criterion

(2), we have

$$\begin{aligned}\|g_m - g_n\|_{\mathcal{H}_0} &= \|g_m - g_n\|_{\mathcal{H}} \\ &\leq \|g_m - f_m\|_{\mathcal{H}} + \|f_m - f_n\|_{\mathcal{H}} + \|f_n - g_n\|_{\mathcal{H}} \\ &\leq \frac{1}{m} + \frac{1}{n} + \|f_m - f_n\|_{\mathcal{H}}\,,\end{aligned}$$

hence $\{g_n\}$ is Cauchy in $\mathcal{H}_0$.

Finally, is this limiting f a limit with respect to the $\mathcal{H}$-norm? Yes, since by Lemma (39) (denseness of $\mathcal{H}_0$ in $\mathcal{H}$: see the first lines of the proof), g_n tends to f in the $\mathcal{H}$-norm sense, and thus f_n converges to f in $\mathcal{H}$-norm,

$$\begin{aligned}\|f_n - f\|_{\mathcal{H}} &\leq \|f_n - g_n\|_{\mathcal{H}} + \|g_n - f\|_{\mathcal{H}} \\ &\leq \frac{1}{n} + \|g_n - f\|_{\mathcal{H}}\,.\end{aligned}$$

Thus $\mathcal{H}$ is complete.

4.5 How to build a valid pre-RKHS $\mathcal{H}_0$

Here we show how to build a valid pre-RKHS. Imporantly, in doing this, we prove that for every positive definite kernel, there corresponds a unique RKHS $\mathcal{H}$.

Theorem 42. (Moore-Aronszajn)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be positive definite. There is a unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel k. Moreover, if space $\mathcal{H}_0 = \left[\{k(\cdot, x)\}_{x \in \mathcal{X}}\right]$ is endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(y_j, x_i), \tag{4.4}$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$, then $\mathcal{H}_0$ is a valid pre-RKHS.

We first need to show that (4.4) is a **valid inner product**. First, is it independent of the particular α_i and β_i used to define f, g? Yes, since

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(y_j).$$

As a useful consequence of this result we get the **reproducing property** on $\mathcal{H}_0$, by setting $g = k(x, \cdot)$,

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x).$$

Next, we check that the form (4.4) is indeed a valid inner product on $\mathcal{H}_0$. The only nontrivial axiom to be verified is

$$\langle f, f\rangle_{\mathcal{H}_0} = 0 \implies f = 0.$$

This is true since

$$\forall x \in \mathcal{X},\ f(x) = \langle f(\cdot), k(x,\cdot)\rangle \underset{(a)}{\leq} \|f\|_{\mathcal{H}_0} k^{1/2}(x,x) = 0,$$

where in (a) we use Cauchy-Schwarz. We now proceed to the main proof.

Proof. (that $\mathcal{H}_0$ satisfies the pre-RKHS axioms). Let $t \in E$. Note that for $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$

$$\langle f, k(\cdot,t)\rangle_{\mathcal{H}_0} = \sum_{i=1}^n \alpha_i k(t, x_i) = f(t), \tag{4.5}$$

and thus for $f, g \in \mathcal{H}_0$,

$$\begin{aligned} |\delta_x(f) - \delta_x(g)| &= |\langle f - g, k(\cdot,x)\rangle_{H_0}| \\ &\leq k^{1/2}(x,x)\,\|f-g\|_{\mathcal{H}_0}, \end{aligned}$$

meaning δ_x is continuous on $\mathcal{H}_0$, and the **first** pre-RKHS requirement is satisfied.

Now, take $\epsilon > 0$ and define a Cauchy $\{f_n\}$ in $\mathcal{H}_0$ that converges pointwise to 0. Since Cauchy sequences are bounded, we may define $A > 0$, s.t. $\|f_n\|_{\mathcal{H}_0} < A$, $\forall n \in \mathbb{N}$. One can find $N_1 \in \mathbb{N}$, s.t. $\|f_n - f_m\|_{H_0} < \epsilon/2A$, for $n, m \geq N_1$. Write $f_{N_1} = \sum_{i=1}^k \alpha_i k(\cdot, x_i)$. Take $N_2 \in \mathbb{N}$, s.t. $|f_n(x_i)| < \frac{\epsilon}{2k|\alpha_i|}$, for $i = 1, \ldots, k$. Now, for $n \geq \max(N_1, N_2)$

$$\begin{aligned} \|f_n\|^2_{\mathcal{H}_0} &\leq |\langle f_n - f_{N_1}, f_n\rangle_{\mathcal{H}_0}| + |\langle f_{N_1}, f_n\rangle_{\mathcal{H}_0}| \\ &\leq \|f_n - f_{N_1}\|_{\mathcal{H}_0}\|f_n\|_{\mathcal{H}_0} + \sum_{i=1}^k |\alpha_i f_n(x_i)| \\ &< \epsilon, \end{aligned}$$

so f_n converges to 0 in $\|\cdot\|_{\mathcal{H}_0}$. Thus, all the pre-RKHS axoims are satisfied, and $\mathcal{H}$ is an RKHS.

To see that the **reproducing kernel** on $\mathcal{H}$ is k, simply note that if $f \in \mathcal{H}$, and $\{f_n\}$ in $\mathcal{H}_0$ converges to f pointwise,

$$\begin{aligned} \langle f, k(\cdot,x)\rangle_{\mathcal{H}} &\underset{(a)}{=} \lim_{n\to\infty} \langle f_n, k(\cdot,x)\rangle_{\mathcal{H}_0} \\ &= \lim_{n\to\infty} f_n(x) \\ &= f(x). \end{aligned}$$

where in (a) we use the defnition of an inner product on $\mathcal{H}$ in (4.2). Since $\mathcal{H}_0$ is dense in $\mathcal{H}$, $\mathcal{H}$ is the unique RKHS that contains $\mathcal{H}_0$. But since $k(\cdot, x) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, it is clear that any RKHS with reproducing kernel k must contain $\mathcal{H}_0$. □

4.6 Summary

Moore-Aronszajn theorem tells us that every positive definite function is a reproducing kernel. We have previously seen that every reproducing kernel is a kernel and that every kernel is a positive definite function. Therefore, all three notions are exactly the same! In addition, we have established a bijection between the set of all positive definite functions on $\mathcal{X} \times \mathcal{X}$, denoted by $\mathbb{R}_+^{\mathcal{X} \times \mathcal{X}}$, and the set of all reproducing kernel Hilbert spaces, denoted by $Hilb(\mathbb{R}^{\mathcal{X}})$, which consists of subspaces of $\mathbb{R}^{\mathcal{X}}$. It turns out that this bijection also preserves the geometric structure of these sets, which are in both cases *closed convex cones*, and we will give some intuition on this in the next Section.

5 Operations with kernels

Since kernels are just positive definite functions, the following Lemma is immediate:

Lemma 43. (Sum and scaling of kernels)
If k, k_1, and k_2 are kernels on $\mathcal{X}$, and $\alpha \geq 0$ is a scalar, then αk, $k_1 + k_2$ are kernels.

Note that a difference of kernels is not necessarily a kernel! This is because we cannot have $k_1(x, x) - k_2(x, x) < 0$, since we would then have a feature map $\phi : \mathcal{X} \to \mathcal{H}$ for which $\langle \phi(x), \phi(x) \rangle_{\mathcal{H}} < 0$. Mathematically speaking, these properties give the set of all kernels the structure of a convex cone (*not* a linear space). Now, consider the following: since we know that $k = k_1 + k_2$ also has an RKHS $\mathcal{H}_k$, what is the relationship between $\mathcal{H}_k$ and the RKHSs $\mathcal{H}_{k_1}$ and $\mathcal{H}_{k_2}$ of k_1 and k_2? The following theorem gives an answer.

Theorem 44. (Sum of RKHSs)
Let $k_1, k_2 \in \mathbb{R}_+^{\mathcal{X} \times \mathcal{X}}$, and $k = k_1 + k_2$. Then,

$$\mathcal{H}_k = \mathcal{H}_{k_1} + \mathcal{H}_{k_2} = \{ f_1 + f_2 \,:\, f_1 \in \mathcal{H}_{k_1}, f_2 \in \mathcal{H}_{k_2} \}, \tag{5.1}$$

and $\forall f \in \mathcal{H}_k$,

$$\|f\|_{\mathcal{H}_k}^2 = \min_{f_1 + f_2 = f} \left\{ \|f_1\|_{\mathcal{H}_{k_1}}^2 + \|f_2\|_{\mathcal{H}_{k_2}}^2 \right\}. \tag{5.2}$$

The product of kernels is also a kernel. Note that this contains as a consequence a familiar fact from linear algebra: that the Hadamard product of two positive definite matrices is positive definite.

Theorem 45. (Product of kernels)
Let k_1 and k_2 be kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then

$$k\left((x,y),(x',y')\right) \quad := \quad k_1(x,x')k_2(y,y')$$

is a kernel on $\mathcal{X}\times\mathcal{Y}$. In addition, there is an isometric isomorphism between $\mathcal{H}_k$ and the Hilbert space tensor product $\mathcal{H}_{k_\mathcal{X}}\otimes\mathcal{H}_{k_\mathcal{Y}}$. In addition, if $\mathcal{X}=\mathcal{Y}$,

$$k\left(x,x'\right) \quad := \quad k_1(x,x')k_2(x,x')$$

is a kernel on $\mathcal{X}$.

The above results enable us to construct many interesting kernels by multiplication, addition and scaling by non-negative scalars. To illustrate this assume that $\mathcal{X}=\mathbb{R}$. The trivial (linear) kernel on $\mathbb{R}^d$ is $k_{lin}(x,x')=\langle x,x'\rangle$. Then for any polynomial $p(t)=a_mt^m+\cdots+a_1t+a_0$ with non-negative coefficients a_i, $p(\langle x,x'\rangle)$ defines a valid kernel on $\mathbb{R}^d$. This gives rise to the **polynomial kernel** $k_{poly}(x,x')=(\langle x,x'\rangle+c)^m$, for $c\geq 0$. One can extend the same argument to all functions which have the Taylor series with non-negative coefficients (Steinwart & Christmann, 2008, Lemma 4.8). This leads us to the **exponential kernel** $k_{exp}(x,x')=\exp(2\sigma\langle x,x'\rangle)$, for $\sigma>0$. Furthermore, let $\phi:\mathbb{R}^d\to\mathbb{R}$, $\phi(x)=\exp(-\sigma\|x\|^2)$. Then, since it is representable as an inner product in $\mathbb{R}$ (i.e., ordinary product), $\tilde{k}(x,x')=\phi(x)\phi(x')=\exp(-\sigma\|x\|^2)\exp(-\sigma\|x'\|^2)$ is a kernel on $\mathbb{R}^d$. Therefore, by Theorem 45, so is:

$$\begin{aligned} k_{gauss}(x,x') &= \tilde{k}(x,x')k_{exp}(x,x') \\ &= \exp\left(-\sigma\left[\|x\|^2+\|x'\|^2-2\langle x,x'\rangle\right]\right) \\ &= \exp\left(-\sigma\|x-x'\|^2\right), \end{aligned}$$

which is the **gaussian kernel** on $\mathbb{R}^d$.

6 Mercer representation of RKHS

Moore-Aronszajn theorem gives a construction of an RKHS without imposing any additional assumptions on $\mathcal{X}$ (apart from it being a non-empty set) nor on kernel k (apart from it being a positive definite function). In this Section, we will consider $\mathcal{X}$ to be a compact metric space (with metric $d_\mathcal{X}$) and that $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a *continuous* positive definite function. These assumptions will allow us to give an alternative construction and interpretation of RKHS.

6.1 Integral operator of a kernel

Definition 46 (Integral operator)**.** Let k be a continuous kernel on compact metric space $\mathcal{X}$, and let ν be a finite Borel measure on $\mathcal{X}$. Let S_k be the linear

map:

$$\begin{aligned} S_k : L_2(\mathcal{X};\nu) &\rightarrow \mathcal{C}(\mathcal{X}), \\ (S_k f)(x) &= \int k(x,y) f(y) d\nu(y), \quad f \in L_2(\mathcal{X};\nu), \end{aligned}$$

and $T_k = I_k \circ S_k$ its composition with the inclusion $I_k : \mathcal{C}(\mathcal{X}) \hookrightarrow L_2(\mathcal{X};\nu)$. T_k is said to be the *integral operator* of kernel k.

Let us first show that the operator S_k is well-defined, i.e., that $S_k f$ is a continuous function $\forall f \in L_2(\mathcal{X};\nu)$. Indeed, $\forall x, y \in \mathcal{X}$, we have that:

$$\begin{aligned} |(S_K f)(x) - (S_K f)(y)| &= \left| \int (k(x,z) - k(y,z)) f(z) d\nu(z) \right| \\ &= \left| \langle k(x,\cdot) - k(y,\cdot), f \rangle_{L_2} \right| \\ &\leq \|k(x,\cdot) - k(y,\cdot)\|_2 \|f\|_2 \\ &\leq \left[\int (k(x,z) - k(y,z))^2 d\nu(z) \right]^{1/2} \|f\|_2 \\ &\leq \sqrt{\nu(\mathcal{X})} \max_{z \in \mathcal{X}} |k(x,z) - k(y,z)| \, \|f\|_2 . \end{aligned}$$

At this point, we use the fact that k is uniformly continuous on $\mathcal{X} \times \mathcal{X}$ (as it is a continuous function on a compact domain. Namely, $\forall \epsilon > 0$, $\exists \delta = \delta(\epsilon)$, s.t. $d_{\mathcal{X}}(x,y) < \delta$ implies $|k(x,z) - k(y,z)| < \frac{\epsilon}{\sqrt{\nu(\mathcal{X})}\|f\|_2}$, $\forall x, y, z \in \mathcal{X}$. From here,

$$d_{\mathcal{X}}(x,y) < \delta \quad \Rightarrow \quad |(S_K f)(x) - (S_K f)(y)| < \epsilon, \qquad \forall x, y \in \mathcal{X},$$

i.e., $S_k f$ is a continuous function on $\mathcal{X}$.

Note that the operator $T_k : L_2(\mathcal{X};\nu) \rightarrow L_2(\mathcal{X};\nu)$ is distinct from S_k. In particular, while $S_k f$ is a continuous function, $T_k f$ is an equivalence class, so $(S_k f)(x)$ is defined, while $(T_k f)(x)$ is *not*.

The integral operator inherits various properties of the kernel function. In particular, it is readily shown that symmetry of k implies that T_k is a *self-adjoint* operator, i.e., that $\langle f, T_k g \rangle = \langle T_k f, g \rangle$, $\forall f, g \in L_2(\mathcal{X};\nu)$, and that positive definiteness of k implies that T_k is a positive operator, i.e., that $\langle f, T_k f \rangle \geq 0$ $\forall f \in L_2(\mathcal{X};\nu)$. Furthermore, continuity of k implies that T_k is also a *compact operator* - the proof of this requires the use of *Arzela-Ascoli theorem*, which can be found in Rudin (1987, Theorem 11.28, p.245). Thus, one can apply an important result of functional analysis, the spectral theorem, to the operator T_k, which states that any compact, self-adjoint operator can be diagonalized in an appropriate orthonormal basis.

Theorem 47. (Spectral theorem)

Let $\mathcal{F}$ be a Hilbert space, and $T : \mathcal{F} \rightarrow \mathcal{F}$ a compact, self-adjoint operator. There is an at most countable orthonormal set (ONS) $\{e_j\}_{j \in J}$ of $\mathcal{F}$ and $\{\lambda_j\}_{j \in J}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ converging to zero, such that

$$Tf = \sum_{j \in J} \lambda_j \langle f, e_j \rangle_{\mathcal{F}} e_j, \qquad f \in \mathcal{F}.$$

6.2 Mercer's theorem

Let us know fix a finite measure ν on $\mathcal{X}$ with $supp\nu = \mathcal{X}$. Recall that the integral operator T_k is compact, positive and self-adjoint on $L_2(\mathcal{X};\nu)$, so, by the spectral theorem, there exists ONS $\{\tilde{e}_j\}_{j\in J}$ and the set of eigenvalues $\{\lambda_j\}_{j\in J}$, where J is at most countable set of indices, corresponding to the *strictly positive eigenvalues* of T_k. Note that each $\tilde{e}_j$ is an equivalence class in the ONS of $L_2(\mathcal{X};\nu)$, but to each equivalence class we can also assign a continuous function $e_j = \lambda_j^{-1} S_k \tilde{e}_j \in \mathcal{C}(\mathcal{X})$. To show that e_j is in the class $\tilde{e}_j$, note that:

$$I_k e_j = \lambda_j^{-1} T_k \tilde{e}_j = \lambda_j^{-1} \lambda_j \tilde{e}_j = \tilde{e}_j.$$

With this notation, the following Theorem holds:

Theorem 48. (Mercer's theorem)
Let k be a continuous kernel on compact metric space $\mathcal{X}$, and let ν be a finite Borel measure on $\mathcal{X}$ with $supp\nu = \mathcal{X}$. Then $\forall x, y \in \mathcal{X}$

$$k(x,y) = \sum_{j\in J} \lambda_j e_j(x) e_j(y),$$

and the convergence of the sum is uniform on $\mathcal{X}\times\mathcal{X}$, and absolute for each pair $(x,y) \in \mathcal{X}\times\mathcal{X}$.

Note that Mercer's theorem gives us another feature map for the kernel k, since:

$$\begin{aligned} k(x,y) &= \sum_{j\in J} \lambda_j e_j(x) e_j(y) \\ &= \left\langle \sqrt{\lambda_j} e_j(x), \sqrt{\lambda_j} e_j(y) \right\rangle_{\ell^2(J)}, \end{aligned}$$

so we can take $\ell^2(J)$ as a feature space, and the corresponding feature map is:

$$\begin{aligned} \phi : \mathcal{X} &\to \ell^2(J) \\ \phi : x &\mapsto \left\{ \sqrt{\lambda_j} e_j(x) \right\}_{j\in J}. \end{aligned}$$

This map is well defined as $\sum_{j\in J} \left|\sqrt{\lambda_j} e_j(x)\right|^2 = k(x,x) < \infty$.

Apart from the representation of the kernel function, Mercer theorem also leads to a construction of RKHS using the eigenfunctions of the integral operator T_k. In particular, first note that sum $\sum_{j\in J} a_j e_j(x)$ converges absolutely $\forall x \in \mathcal{X}$ whenever sequence $\left\{ \frac{a_j}{\sqrt{\lambda_j}} \right\} \in \ell^2(J)$. Namely, from the Cauchy-Schwartz

inequality in $\ell^2(J)$, we have that:

$$\begin{aligned}\sum_{j\in J}|a_j e_j(x)| &\leq \left[\sum_{j\in J}\left|\frac{a_j}{\sqrt{\lambda_j}}\right|^2\right]^{1/2}\cdot\left[\sum_{j\in J}\left|\sqrt{\lambda_j}e_j(x)\right|^2\right]^{1/2}\\ &= \left\|\left\{\frac{a_j}{\sqrt{\lambda_j}}\right\}\right\|_{\ell^2}\sqrt{k(x,x)}.\end{aligned}$$

In that case, $\sum_{j\in J}a_j e_j$ is a well defined function on $\mathcal{X}$. The following theorem tells us that the RKHS of k is exactly the space of functions of this form.

Theorem 49. *Let $\mathcal{X}$ be a compact metric space and $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ a continuous kernel. Define:*

$$\mathcal{H} = \left\{f=\sum_{j\in J}a_j e_j \;:\; \left\{\frac{a_j}{\sqrt{\lambda_j}}\right\}\in\ell^2(J)\right\},$$

with inner product:

$$\left\langle\sum_{j\in J}a_j e_j,\sum_{j\in J}b_j e_j\right\rangle_{\mathcal{H}} = \sum_{j\in J}\frac{a_j b_j}{\lambda_j}. \tag{6.1}$$

Then $\mathcal{H}=\mathcal{H}_k$ (they are the same spaces of functions with the same inner product).

Proof. Routine work shows that (6.1) defines an inner product and that $\mathcal{H}$ is a Hilbert space. By Mercer's theorem, $k(\cdot,x)=\sum_{j\in J}(\lambda_j e_j(x))\,e_j$, and:

$$\begin{aligned}\sum_{j\in J}\left|\frac{\lambda_j e_j(x)}{\sqrt{\lambda_j}}\right|^2 &= \sum_{j\in J}\lambda_j e_j^2(x)\\ &= k(x,x)<\infty,\end{aligned}$$

so $k(\cdot,x)\in\mathcal{H}$, $\forall x\in\mathcal{X}$. Furtheromre, let $f=\sum_{j\in J}a_j e_j\in\mathcal{H}$ with $\left\{\frac{a_j}{\sqrt{\lambda_j}}\right\}\in\ell^2(J)$. Then,

$$\begin{aligned}\langle f,k(\cdot,x)\rangle_{\mathcal{H}} &= \left\langle\sum_{j\in J}a_j e_j,\sum_{j\in J}(\lambda_j e_j(x))\,e_j\right\rangle_{\mathcal{H}}\\ &= \sum_{j\in J}\frac{a_j\lambda_j e_j(x)}{\lambda_j}\\ &= f(x).\end{aligned}$$

Thus, $\mathcal{H}$ is a Hilbert space of functions with a reproducing kernel k, so it must be equal to $\mathcal{H}_k$ by the uniqueness of RKHS. □

A consequence of the above theorem is that although space $\mathcal{H}$ is defined using the integral operator T_k and its associated eigenfunctions $\{e_j\}_{j\in J}$ which depend on the underlying measure ν, it coincides exactly with the RKHS $\mathcal{H}_k$ of k, so by uniqueness of the RKHS shown in the Moore-Aronszajn theorem, $\mathcal{H}$ actually does not depend on the choice of ν at all.

6.3 Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

Assume now that $\{\tilde{e}_j\}_{j\in J}$ is an orthonormal basis of $L_2(\mathcal{X};\nu)$, i.e., that all eigan-values of T_k are strictly positive. Write $\hat{f}(j) = \langle f, \tilde{e}_j\rangle_{L_2}$ for Fourier coefficients of $f \in L_2(\mathcal{X};\nu)$, w.r.t. the basis $\{\tilde{e}_j\}_{j\in J}$. Then,

$$T_k f = \sum_{j\in J} \lambda_j \hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu),$$

so in the expansion w.r.t. orthonormal basis $\{\tilde{e}_j\}_{j\in J}$, T_k simply scales the Fourier coefficients with respective eigenvalues. The operator $T_k^{1/2}$ for which $T_k = T_k^{1/2} \circ T_k^{1/2}$ is given by:

$$T_k^{1/2} f = \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu).$$

Note that if we replace classes $\tilde{e}_j$ with their representers $e_j = \lambda_j^{-1} S_k \tilde{e}_j$, we obtain a function in the RKHS, i.e.,

$$\sum_{j\in J} \left|\hat{f}(j)\right|^2 = \|f\|_2^2 < \infty \Rightarrow \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \Rightarrow \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

Thus, $T_k^{1/2}$ induces an isometric isomorphism between $L_2(\mathcal{X};\nu)$ and $\mathcal{H}_k$ (and both are isometrically isomorphic to $\ell^2(J)$). In the case where not all eigenvalues of T_k are strictly positive, $\{\tilde{e}_j\}_{j\in J}$ does not span all of $L_2(\mathcal{X};\nu)$, but $\mathcal{H}_k$ is still isometrically isomorphic to its subspace $span\,\{\tilde{e}_j \ : j \in J\} \subseteq L_2(\mathcal{X};\nu)$.

7 Further results

- Separable RKHS: Steinwart & Christmann (2008, Lemma 4.33)
- Measurability of canonical feature map: Steinwart & Christmann (2008, Lemma 4.25)
- Relation between RKHS and $L_2(\mu)$: Steinwart & Christmann (2008, Theorem 4.26, Theorem 4.27). Note in particular Steinwart & Christmann (2008, Theorem 4.47): the mapping from L_2 to $\mathcal{H}$ for the Gaussian RKHS is injective.

- Expansion of kernel in terms of basis functions: Berlinet & Thomas-Agnan (2004, Theorem 14 p. 32)
- Mercer's theorem: Steinwart & Christmann (2008, p. 150).

8 What functions are in an RKHS?

- Gaussian RKHSs do not contain constants: Steinwart & Christmann (2008, Corollary 4.44).
- Universal RKHSs are dense in the space of bounded continuous functions: Steinwart & Christmann (2008, Section 4.6)
- The bandwidth of the kernel limits the bandwidth of the functions in the RKHS: (c.f., e.g., Appendix of Christian Walder PhD thesis, University of Queensland, 2008).

9 Acknowledgements

Thanks to Sivaraman Balakrishnan for careful proofreading.

References

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer, 2004.

Kreyszig, E. *Introductory Functional Analysis with Applications.* Wiley, 1989.

Rudin, W. *Real and Complex Analysis.* McGraw-Hill, 3rd edition, 1987.

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines.* Information Science and Statistics. Springer, 2008.

Wendland, H. *Scattered Data Approximation.* Cambridge University Press, Cambridge, UK, 2005.