Lecture 1: Introduction to RKHS
Columbia, 2014

Gatsby Unit, CSML, UCL

April 28, 2014
Kernels and feature space (1): XOR example

- No linear classifier separates red from blue
- Map points to higher dimensional feature space:
  \[ \phi(x) = \begin{bmatrix} x_1 & x_2 & x_1 x_2 \end{bmatrix} \in \mathbb{R}^3 \]
Kernel methods can control **smoothness** and avoid overfitting/underfitting.
We will describe in order:

1. Hilbert space
2. Kernel (lots of examples: e.g. you can build kernels from simpler kernels)
3. Reproducing property
**Hilbert space**

**Definition (Inner product)**

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is an **inner product** on $\mathcal{H}$ if

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

**Norm** induced by the inner product: $\|f\|_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$

**Definition (Hilbert space)**

Inner product space containing Cauchy sequence limits.
Definition (Cauchy sequence)

A sequence \( \{f_n\}_{n=1}^{\infty} \) of elements of a normed vector space \((F, \|\cdot\|_F)\) is said to be a Cauchy (fundamental) sequence if for every \( \varepsilon > 0 \), there exists \( N = N(\varepsilon) \in \mathbb{N} \), such that for all \( n, m \geq N \),
\[
\|f_n - f_m\|_F < \varepsilon.
\]

Definition (Complete space)

A metric space \( F \) is said to be complete if every Cauchy sequence \( \{f_n\}_{n=1}^{\infty} \) in \( F \) converges: it has a limit, and this limit is in \( F \).

- Complete + norm = **Banach space**
- Complete + inner product = **Hilbert space**
Kernel

**Definition**

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

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

- Almost no conditions on $\mathcal{X}$ (e.g., $\mathcal{X}$ itself doesn't need an inner product, e.g., documents).
- A single kernel can correspond to several possible features. A trivial example for $\mathcal{X} := \mathbb{R}:

$$\phi_1(x) = x \quad \text{and} \quad \phi_2(x) = \begin{bmatrix} x/\sqrt{2} \\ x/\sqrt{2} \end{bmatrix}$$
New kernels from old: sums, transformations

Theorem (Sums of kernels are kernels)

Given $\alpha > 0$ and $k$, $k_1$ and $k_2$ all kernels on $\mathcal{X}$, then $\alpha k$ and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove this, just check inner product definition. A difference of kernels may not be a kernel (why?)

Theorem (Mappings between spaces)

Let $\mathcal{X}$ and $\tilde{\mathcal{X}}$ be sets, and define a map $A : \mathcal{X} \to \tilde{\mathcal{X}}$. Define the kernel $k$ on $\tilde{\mathcal{X}}$. Then the kernel $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

Example: $k(x, x') = x^2 (x')^2$. 
New kernels from old: products

Theorem (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.

Define:

- $\mathcal{H}_1$ corresponding to $k_1(x, x') = \langle \phi_1(x), \phi_1(x') \rangle_{\mathcal{H}_1}$ (e.g.: kernel between two images)
- $\mathcal{H}_2$ corresponding to $k_2(y, y') = \langle \phi_2(y), \phi_2(y') \rangle_{\mathcal{H}_2}$ (e.g.: kernel between two captions)

Is the following a kernel? (e.g. between one image-caption pair and another)

$$K \left[ (x, y), (x', y') \right] = k_1(x, x') \times k_2(y, y')$$
New kernels from old: products

Proof.

Given $b \in \mathcal{H}_2$ and $a \in \mathcal{H}_1$, we define the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{H}_2$ to $\mathcal{H}_1$,

$$(a \otimes b)f \mapsto \langle b, f \rangle_{\mathcal{H}_2} a.$$

$a \otimes b \in \text{HS}(\mathcal{H}_2, \mathcal{H}_1)$, space of Hilbert-Schmidt operators with inner product

$$\langle L, M \rangle_{\text{HS}} = \sum_{j \in J} \langle Lf_j, Mf_j \rangle_{\mathcal{H}_1},$$

where $\{f_j\}$ an ONB of $\mathcal{H}_2$. Applying the above definition, for any $L \in \text{HS}(\mathcal{G}, \mathcal{F})$,

$$\langle L, a \otimes b \rangle_{\text{HS}} = \langle a, Lb \rangle_{\mathcal{H}_1}.$$
Proof.

To see this, first expand \( b = \sum_{j \in J} \langle b, f_j \rangle_{\mathcal{H}_2} f_j \). Then

\[
\langle a, Lb \rangle = \left\langle a, L \left( \sum_j \langle b, f_j \rangle_{\mathcal{H}_2} f_j \right) \right\rangle_{\mathcal{H}_1} \\
= \sum_j \langle b, f_j \rangle_{\mathcal{H}_2} \langle a, Lf_j \rangle_{\mathcal{H}_1}
\]

and

\[
\langle a \otimes b, L \rangle_{\text{HS}} = \sum_j \langle Lf_j, (a \otimes b)f_j \rangle_{\mathcal{H}_1} \\
= \sum_j \langle b, f_j \rangle_{\mathcal{H}_2} \langle a, Lf_j \rangle_{\mathcal{H}_1}.
\]
New kernels from old: products

Proof.

Special case:

\[ \langle u \otimes v, a \otimes b \rangle_{\text{HS}} = \langle u, a \rangle_{\mathcal{H}_1} \langle b, v \rangle_{\mathcal{H}_2}. \]

Apply this to

\[ k_1(x, x')k_2(y, y') = \langle \phi_1(x), \phi_1(x') \rangle_{\mathcal{H}_1} \langle \phi_2(y), \phi_2(y') \rangle_{\mathcal{H}_2} \]

\[ = \langle \phi_1(x) \otimes \phi_2(y), \phi_1(x') \otimes \phi_2(y') \rangle_{\text{HS}}. \]
Theorem (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

\[
k(x, x') := (\langle x, x' \rangle + c)^m
\]

is a valid kernel.

To prove: expand 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.
The kernels we’ve seen so far are dot products between finitely many features. E.g.

\[ k(x, y) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix}^\top \begin{bmatrix} \sin(y) & y^3 & \log y \end{bmatrix} \]

where \( \phi(x) = \begin{bmatrix} \sin(x) & x^3 & \log x \end{bmatrix} \)

Can a kernel be a dot product between infinitely many features?
Infinite sequences

Definition

The space $\ell_p$ 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 $\ell_2$.

Theorem

Given sequence of functions $(\phi_i(x))_{i \geq 1}$ in $\ell_2$ where $\phi_i : \mathcal{X} \to \mathbb{R}$ is the $i$th coordinate of $\phi(x)$. Then

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

(1)
Proof: We just need to check that inner product remains finite. Norm $\|a\|_{\ell_2}$ associated with inner product (1)

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

where $a$ represents sequence with terms $a_i$. Via Cauchy-Schwarz,

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

so the sequence defining the inner product converges for all $x, x' \in \mathcal{X}$.
Taylor series kernels

Definition (Taylor series kernel)

For $r \in (0, \infty]$, with $a_n \geq 0$ for all $n \geq 0$

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

Define $\mathcal{X}$ to be the $\sqrt{r}$-ball in $\mathbb{R}^d$, so $\|x\| < \sqrt{r}$,

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

Example (Exponential kernel)

$$k(x, x') := \exp(\langle x, x' \rangle).$$
Taylor series kernel (proof)

Proof: By Cauchy-Schwarz,

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

so the Taylor series converges. Define $c_{j_1 \ldots j_d} = \frac{n!}{\prod_{i=1}^{d} j_i !}$

$$k(x, x') = \sum_{n=0}^{\infty} a_n \left( \sum_{j=1}^{d} x_j x'_j \right)^n$$

$$= \sum_{n=0}^{\infty} a_n \sum_{j_1 \ldots j_d \geq 0, \ j_1 + \ldots + j_d = n} c_{j_1 \ldots j_d} \prod_{i=1}^{d} (x_i, x'_i)^{j_i}$$

$$= \sum_{j_1 \ldots j_d > 0} a_{j_1 + \ldots + j_d} c_{j_1 \ldots j_d} \prod_{i=1}^{d} x_i^{j_i} \prod_{i=1}^{d} (x'_i)^{j_i}.$$
Example (Gaussian kernel)

The Gaussian kernel on $\mathbb{R}^d$ is defined as

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

**Proof**: an exercise! Use product rule, mapping rule, exponential kernel.
If we are given a function of two arguments, \( k(x, x') \), how can we determine if it is a valid kernel?

1. Find a feature map?
   - Sometimes this is not obvious (e.g. if the feature vector is infinite dimensional, e.g. the Gaussian kernel in the last slide)
   - The feature map is not unique.

2. A direct property of the function: positive definiteness.
Definition (Positive definite functions)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is positive definite if
\[ \forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n, \]
\[ \sum_{i=1}^{n} \sum_{j=1}^{n} a_ia_jk(x_i, x_j) \geq 0. \]

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.
Theorem

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} =: k(x,y)$ is positive definite.

Proof.

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

$$= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.$$

Reverse also holds: positive definite $k(x, x')$ is inner product in a unique $\mathcal{H}$ (Moore-Aronsajn: coming later!).
The reproducing kernel Hilbert space
First example: finite space, polynomial features

Reminder: XOR example:
First example: finite space, polynomial features

Reminder: Feature space from XOR motivating example:

\[ \phi : \mathbb{R}^2 \to \mathbb{R}^3 \]

\[ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \phi(x) = \begin{bmatrix} x_1 \\ x_2 \\ x_1x_2 \end{bmatrix}, \]

with kernel

\[ k(x, y) = \begin{bmatrix} x_1 \\ x_2 \\ x_1x_2 \end{bmatrix}^\top \begin{bmatrix} y_1 \\ y_2 \\ y_1y_2 \end{bmatrix} \]

(the standard inner product in \( \mathbb{R}^3 \) between features). Denote this feature space by \( \mathcal{H} \).
First example: finite space, polynomial features

Define a **linear function** of the inputs $x_1, x_2$, and their product $x_1 x_2$,

$$f(x) = f_1 x_1 + f_2 x_2 + f_3 x_1 x_2.$$  

$f$ in a space of functions mapping from $\mathcal{X} = \mathbb{R}^2$ to $\mathbb{R}$. Equivalent representation for $f$,

$$f(\cdot) = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix}^T.$$  

$f(\cdot)$ refers to the function as an object (here as a **vector** in $\mathbb{R}^3$) 

$f(x) \in \mathbb{R}$ is function evaluated at a point (a **real number**).

$$f(x) = f(\cdot)^\top \phi(x) = \langle f(\cdot), \phi(x) \rangle_\mathcal{H}$$

Evaluation of $f$ at $x$ is an **inner product in feature space** (here standard inner product in $\mathbb{R}^3$) 

$\mathcal{H}$ is a space of functions mapping $\mathbb{R}^2$ to $\mathbb{R}$. 

---

**Lecture 1: Introduction to RKHS**
First example: finite space, polynomial features

φ(y) is a mapping from \( \mathbb{R}^2 \) to \( \mathbb{R}^3 \)...

...which also parametrizes a function mapping \( \mathbb{R}^2 \) to \( \mathbb{R} \).

\[
k(\cdot, y) := \begin{bmatrix} y_1 & y_2 & y_1y_2 \end{bmatrix}^\top = \phi(y),
\]

Given \( y \), there is 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 \( c = y_1y_2 \)

Due to symmetry,

\[
\langle k(\cdot, x), \phi(y) \rangle = uy_1 + vy_2 + wy_1y_2
\]

\[
= k(x, y).
\]

We can write \( \phi(x) = k(\cdot, x) \) and \( \phi(y) = k(\cdot, y) \) without ambiguity:

canonical feature map
This example illustrates the two defining features of an RKHS:

- **The reproducing property:**
  \[
  \forall x \in \mathcal{X}, \forall f(\cdot) \in \mathcal{H}, \quad \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)
  \]
  
  ...or use shorter notation \( \langle f, \phi(x) \rangle_{\mathcal{H}} \).

- In particular, for any \( x, y \in \mathcal{X} \),
  \[
  k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}.
  \]

Note: the feature map of every point is in the feature space:

\[
\forall x \in \mathcal{X}, \quad k(\cdot, x) = \phi(x) \in \mathcal{H},
\]
First example: finite space, polynomial features

Another, more subtle point: $\mathcal{H}$ can be larger than all $\phi(x)$. Why?

E.g. $f = [1 \ 1 \ -1] \in \mathcal{H}$ cannot be obtained by $\phi(x) = [x_1 \ x_2 \ (x_1 x_2)]$. 
Second (infinite) example: Fourier series

Function on the torus $\mathbb{T} := [-\pi, \pi]$ with periodic boundary. Fourier series:

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(i\ell x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell (\cos(\ell x) + i \sin(\ell x)).$$

Example: “top hat” function,

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

Fourier series:

$$\hat{f}_\ell := \frac{\sin(\ell T)}{\ell \pi} \quad f(x) = \sum_{\ell=0}^{\infty} 2\hat{f}_\ell \cos(\ell x).$$
Fourier series for top hat function
Fourier series for top hat function

- \( f(x) \) vs. \( x \)
- Basis function: \( \cos(\ell \times x) \)
- Fourier series coefficients: \( f_{\ell} \)

**Lecture 1: Introduction to RKHS**
Fourier series for top hat function

Top hat

Basis function

Fourier series coefficients

Lecture 1: Introduction to RKHS
Fourier series for top hat function

- Top hat function
- Basis function
- Fourier series coefficients

\[ f(x) = \text{Top hat} \]

\[ \cos(\ell x) \]

\[ f_\ell \]
Fourier series for kernel function

Kernel takes a single argument,

\[ k(x, y) = k(x - y), \]

Define the Fourier series representation of \( k \)

\[ k(x) = \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(i\ell x), \]

\( k \) and its Fourier transform are real and symmetric. E.g. Gaussian,

\[ k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad \hat{k}_\ell \approx \frac{1}{2\pi} \exp\left(-\frac{\sigma^2 \ell^2}{2}\right). \]

Warning: representation \( \hat{k}_\ell \) is an approximation, assumes Gaussian sufficiently narrower than \([-\pi, \pi]\).
Fourier series for Gaussian kernel

Gaussian kernel $k(x)$ compared to the basis function $\cos(\ell \times x)$ for a Fourier series. The Fourier series coefficients $f_\ell$ are also shown.
Fourier series for Gaussian kernel

Gaussian

Basis function

Fourier series coefficients

Lecture 1: Introduction to RKHS
Fourier series for Gaussian kernel

- **Gaussian**: The graph shows the Gaussian kernel function $k(x)$ and its corresponding basis function $\cos(\ell \cdot x)$. The kernel values range from approximately -1 to 1, while the basis function oscillates between -1 and 1.

- **Basis function**: The basis function is plotted for different values of $\ell$, ranging from $-10$ to $10$. The Fourier series coefficients $f_\ell$ are also shown, indicating the amplitude of each basis function component.

**Lecture 1: Introduction to RKHS**
Fourier series for Gaussian kernel

Gaussian

\[ k(x) \]

Basis function

\[ \cos(\ell \times x) \]

Fourier series coefficients

\[ \hat{f}_\ell \]

Lecture 1: Introduction to RKHS
Define $\mathcal{H}$ to be the space of functions with (infinite) feature space representation

$$f(\cdot) = \left[ \ldots \frac{\hat{f}_\ell}{\sqrt{\hat{k}_\ell}} \ldots \right]^T.$$  

The space $\mathcal{H}$ has an inner product:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell = -\infty}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\left(\sqrt{\hat{k}_\ell}\right) \left(\sqrt{\hat{k}_\ell}\right)}.$$  

Define the feature map

$$k(\cdot, x) = \phi(x) = \left[ \ldots \sqrt{\hat{k}_\ell} \exp(-\ell x) \ldots \right]^T.$$
The reproducing theorem holds,

\[
\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \sqrt{k_\ell} \exp(-i\ell x) \frac{1}{\sqrt{k_\ell}} \\
= \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(i\ell x) = f(x),
\]

...including for the kernel itself,

\[
\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \left( \sqrt{k_\ell} \exp(-i\ell x) \right) \left( \sqrt{k_\ell} \exp(-i\ell y) \right) \\
= \sum_{\ell=-\infty}^{\infty} \hat{k}_\ell \exp(i\ell(y - x)) = k(x - y).
\]
Fourier series: what does it achieve?

The squared norm of a function \( f \) in \( \mathcal{H} \) is:

\[
\| f \|_\mathcal{H}^2 = \langle f, f \rangle_\mathcal{H} = \sum_{\ell = -\infty}^{\infty} \frac{\hat{f}_\ell \hat{f}_\ell}{\hat{k}_\ell}.
\]

If \( \hat{k}_\ell \) decays fast, then so must \( \hat{f}_\ell \) if we want \( \| f \|_\mathcal{H}^2 < \infty \). Recall

\[
f(x) = \sum_{\ell = -\infty}^{\infty} \hat{f}_\ell (\cos(\ell x) + i \sin(\ell x)).
\]

Enforces smoothness.

Question: is the \textbf{top hat} function in the Gaussian RKHS?
Reproducing property for function with Gaussian kernel:
\[ f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \langle \sum_{i=1}^{m} \alpha_i \phi(x_i), \phi(x) \rangle_{\mathcal{H}}. \]

- What do the features \( \phi(x) \) look like (there are infinitely many of them, they are not unique!)
- What do these features have to do with smoothness?
Third example: infinite feature space

Define RKHS kernel $k$ such that $\|k\|_{L^2(\mu)} < \infty$ and the associated RKHS $\mathcal{H}$ is separable. The operator

$$T_k : L^2(\mu) \rightarrow L^2(\mu)$$

$$f \mapsto \int f(x') k(x, x') d\mu(x')$$

is compact, positive, self-adjoint. (Steinwart and Christmann, Theorem 4.27)

By the spectral theorem there is an at most countable ONS s.t.

$$T_k f = \sum_j \lambda_j \langle f, e_j \rangle e_j \quad \int_{\chi} e_i(x)e_j(x)d\mu(x) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$ 

Can we use the $\{\lambda_i, e_i\}$ to construct a feature space for $\mathcal{H}$?
Theorem

(Mercer) Let $\mathcal{X}$ be a compact metric space, $k$ be a continuous kernel, and $\mu$ be a finite Borel measure with $\text{supp}\{\mu\} = \mathcal{X}$. Then the convergence of

$$k(x, y) = \sum_{j} \lambda_j e_j(x)e_j(y)$$

is absolute and uniform ($e_j$ is the continuous element of the $L^2$ equivalence class $e_j$).
Theorem

(Mercer RKHS) (Steinwart and Christmann, Theorem 4.51) Under the assumptions of Mercer’s theorem,

\[ \mathcal{H} := \left\{ \sum_i a_i \sqrt{\lambda_i} e_i : a_i \in \ell_2 \right\} \]  \hspace{1cm} (2)

is an RKHS with kernel \( k \). The feature map is \( \phi(x) = \left[ \ldots \sqrt{\lambda_i} e_i(x) \ldots \right] \).

Given two functions in the RKHS

\[ f := \sum_i a_i \sqrt{\lambda_i} e_i, \quad g := \sum_i b_i \sqrt{\lambda_i} e_i, \]

the inner product is \( \langle f, g \rangle_\mathcal{H} = \sum_i a_i b_i \).
Proof: Most of the requirements for this being a Hilbert space are straightforward. There are two aspects requiring care:

1. Is \( k(x, \cdot) \in \mathcal{H} \) \( \forall x \in \mathcal{X} \)? Requires Mercer’s theorem
2. Does the reproducing property hold? \( \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \).

First part:
By the definition of \( \mathcal{H} \) (2), the function in \( \mathcal{H} \) indexed by \( x \) is

\[
k(x, \cdot) = \sum_i \left( \sqrt{\lambda_i} e_i(x) \right) \left( \sqrt{\lambda_i} e_i(\cdot) \right).
\]

Is this function in the RKHS? Yes, if the \( \ell_2 \) norm of \( \sqrt{\lambda_i} e_i(x) \) is bounded. This is due to Mercer: \( \forall x \in \mathcal{X} \),

\[
\left\| \left( \sqrt{\lambda_i} e_i(x) \right) \right\|_{\ell_2}^2 = k(x, x) < \infty.
\]
Proof:
Second part: The reproducing property holds: using the inner product definition,

\[ \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = \sum_i f_i \left( \sqrt{\lambda_i} e_i(x) \right) = f(x), \]

which is always well defined since both \( f \in \ell_2 \) and \( k(x, \cdot) \in \ell_2 \).
Third example: infinite feature space

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

$$
\lambda_k \propto b^k \quad b < 1
$$

$$
e_k(x) \propto \exp(- (c-a)x^2)H_k(x\sqrt{2c}),
$$

$a, b, c$ are functions of $\sigma$, and $H_k$ is $k$th order Hermite polynomial.

$$
k(x, x') = \sum_{i=1}^{\infty} \lambda_i e_i(x) e_i(x')
$$

(Figure from Rasmussen and Williams)

WARNING: $\mathbb{R}$ is non-compact domain, cannot use Mercer argument in form given earlier.
Example RKHS function, Gaussian kernel:

\[ f(x) := \sum_{i=1}^{m} \alpha_i k(x_i, x) = \sum_{i=1}^{m} \alpha_i \left[ \sum_{j=1}^{\infty} \lambda_j e_j(x_i) e_j(x) \right] = \sum_{j=1}^{\infty} f_j \left[ \sqrt{\lambda_j} e_j(x) \right] \]

where \( f_j = \sum_{i=1}^{m} \alpha_i \sqrt{\lambda_j} e_j(x_i) \).

**NOTE** that this enforces smoothing:
\( \lambda_j \) decay as \( e_j \) become rougher, \( f_j \) decay since
\( \sum_j f_j^2 < \infty \).
Some reproducing kernel Hilbert space theory
Reproducing kernel Hilbert space (1)

Definition

$\mathcal{H}$ a Hilbert space of $\mathbb{R}$-valued functions on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if

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

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

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_\mathcal{H}. \quad (3)$$

Original definition: kernel an inner product between feature maps. Then $\phi(x) = k(\cdot, x)$ a valid feature map.
Another RKHS definition:
Define $\delta_x$ to be the operator of evaluation at $x$, i.e.

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

Definition (Reproducing kernel Hilbert space)

$\mathcal{H}$ is an RKHS if the evaluation operator $\delta_x$ is bounded: $\forall x \in \mathcal{X}$ there exists $\lambda_x \geq 0$ such that for all $f \in \mathcal{H}$,

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

$\implies$ two functions identical in RHKS norm agree at every point:

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

**Theorem (Reproducing kernel equivalent to bounded $\delta_x$)**

$\mathcal{H}$ is a reproducing kernel Hilbert space (i.e., its evaluation operators $\delta_x$ are bounded linear operators), if and only if $\mathcal{H}$ has a reproducing kernel.

**Proof:** If $\mathcal{H}$ has a reproducing kernel $\iff \delta_x$ bounded

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

Cauchy-Schwarz in 3rd line. Consequently, $\delta_x : \mathcal{F} \to \mathbb{R}$ bounded with $\lambda_x = k(x, x)^{1/2}$. 

Lecture 1: Introduction to RKHS
**Proof:** \( \delta_x \) bounded \( \implies \mathcal{H} \) has a reproducing kernel

We use...

**Theorem**

*(Riesz representation) In a Hilbert space \( \mathcal{H} \), all bounded linear functionals are of the form \( \langle \cdot, g \rangle_{\mathcal{H}} \), for some \( g \in \mathcal{H} \).*

If \( \delta_x : \mathcal{F} \to \mathbb{R} \) is a bounded linear functional, by Riesz \( \exists f_{\delta_x} \in \mathcal{H} \) such that

\[
\delta_x f = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.
\]

Define \( k(x', x) = f_{\delta_x}(x') \), \( \forall x, x' \in \mathcal{X} \). By its definition, both \( k(\cdot, x) = f_{\delta_x} \in \mathcal{H} \) and \( \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \delta_x f = f(x) \). Thus, \( k \) is the reproducing kernel.
Theorem (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 $k$.

Recall feature map is not unique (as we saw earlier): only kernel is.
Main message #1

Reproducing kernels

Positive definite functions

Hilbert function spaces with bounded point evaluation
Small RKHS norm results in smooth functions.

E.g. kernel ridge regression with Gaussian kernel:

\[
\hat{f} = \arg \min_{f \in \mathcal{H}} \left( \sum_{i=1}^{n} (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \| f \|_{\mathcal{H}}^2 \right).
\]
Moore-Aronszajn Theorem: pre-RKHS

How do we prove this? (Sketch only - a very good full proof is in Berlinet and Thomas-Agnan, 2004, Chapter 1)

Starting with a positive def. $k$, construct a pre-RKHS (an inner product space) $\mathcal{H}_0 \subset \mathbb{R}^X$ with properties:

1. The evaluation functionals $\delta_x$ are continuous on $\mathcal{H}_0$,
2. Any $\mathcal{H}_0$-Cauchy sequence $f_n$ which converges pointwise to 0 also converges in $\mathcal{H}_0$-norm to 0
Moore-Aronszajn Theorem: pre-RKHS

pre-RKHS $\mathcal{H}_0 = \text{span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$ will be taken to be the set of functions:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$
Theorem (Moore-Aronszajn - Step A)

Space $\mathcal{H}_0 = \text{span} \{ k(\cdot, x) | x \in \mathcal{X} \}$, endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j),$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$, is a valid pre-RKHS.

Theorem (Moore-Aronszajn - Step B)

Let $\mathcal{H}_0$ be a pre-RKHS space. 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 $\{f_n\}$ converging pointwise to $f$. Then, $\mathcal{H}$ is an RKHS.
Moore-Aronszajn Theorem - Step A

- Is \( \langle f, g \rangle_{\mathcal{H}_0} \) a valid inner product?
- Are evaluation functionals \( \delta_x \) are continuous on \( \mathcal{H}_0 \)?
- Does every \( \mathcal{H}_0 \)-Cauchy sequence \( f_n \) which converges pointwise to 0 also converge in \( \mathcal{H}_0 \)-norm to 0?
Moore-Aronszajn Theorem- Step B

Define $\mathcal{H}$ to be the set of functions $f \in \mathbb{R}^X$ for which there exists an $\mathcal{H}_0$-Cauchy sequence $\{f_n\}$ converging pointwise to $f$. Clearly, $\mathcal{H}_0 \subseteq \mathcal{H}$.

1. We define the inner product between $f, g \in \mathcal{H}$ as the limit of an inner product of the $\mathcal{H}_0$-Cauchy sequences $\{f_n\}, \{g_n\}$ converging to $f$ and $g$ respectively. Is this inner product well defined, i.e., independent of the sequences used?

2. 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?

3. Are the evaluation functionals still continuous on $\mathcal{H}$?

4. Is $\mathcal{H}$ complete (i.e., does every $\mathcal{H}$-Cauchy sequence converge)?

$(1)+(2)+(3)+(4) \implies \mathcal{H} \text{ is RKHS!}$
Kernel Ridge Regression
Kernel ridge regression

Very simple to implement, works well when no outliers.
Ridge regression: case of $\mathbb{R}^D$

We are given $n$ training points in $\mathbb{R}^D$:

$$X = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \in \mathbb{R}^{D \times n} \quad y := \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^\top$$

Define some $\lambda > 0$. Our goal is:

$$f^* = \arg \min_{f \in \mathbb{R}^d} \left( \sum_{i=1}^n (y_i - x_i^\top f)^2 + \lambda \|f\|^2 \right)$$

The second term $\lambda \|f\|^2$ is chosen to avoid problems in high dimensional spaces (see below).
Ridge regression: case of $\mathbb{R}^D$

We are given $n$ training points in $\mathbb{R}^D$:

$$X = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \in \mathbb{R}^{D \times n} \quad y := \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^\top$$

Define some $\lambda > 0$. Our goal is:

$$a^* = \arg \min_{f \in \mathbb{R}^d} \left( \sum_{i=1}^n (y_i - x_i^\top f)^2 + \lambda \| f \|^2 \right)$$

Solution is:

$$f^* = \left( XX^\top + \lambda I \right)^{-1} X y,$$

which is the classic regularized least squares solution.
Kernel ridge regression

Use features of $\phi(x_i)$ in the place of $x_i$:

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

E.g. for finite dimensional feature spaces,

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

$a$ is a vector of length $\ell$ giving weight to each of these features so as to find the mapping between $x$ and $y$. Feature vectors can also have infinite length (more soon).
Kernel ridge regression

Solution easy if we already know \( f \) is a linear combination of feature space mappings of points: representer theorem.

\[
f = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).
\]
Given a set of paired observations \((x_1, y_1), \ldots, (x_n, y_n)\) (regression or classification).
Find the function \(f^*\) in the RKHS \(H\) which satisfies

\[
J(f^*) = \min_{f \in H} J(f),
\]

where

\[
J(f) = L_y(f(x_1), \ldots, f(x_n)) + \Omega \left( \|f\|_H^2 \right),
\]

\(\Omega\) is non-decreasing, and \(y\) is the vector of \(y_i\).

- **Classification:** \(L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^{n} \mathbb{I}_{y_i f(x_i) \leq 0}\)
- **Regression:** \(L_y(f(x_1), \ldots, f(x_n)) = \sum_{i=1}^{n} (y_i - f(x_i))^2\)
The representer theorem: (simple version) solution to

$$
\min_{f \in \mathcal{H}} \left[ L_y(f(x_1), \ldots, f(x_n)) + \Omega \left( \| f \|_{\mathcal{H}}^2 \right) \right]
$$

takes the form

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

If $\Omega$ is strictly increasing, all solutions have this form.
Representer theorem: proof

**Proof:** Denote $f_s$ projection of $f$ onto the subspace

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

such that

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

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

**Regularizer:**

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

then

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

so this term is minimized for $f = f_s$. 
Representer theorem: proof

Proof (cont.): Individual terms $f(x_i)$ in the loss:

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

so

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

Hence

- Loss $L(\ldots)$ only depends on the component of $f$ in the data subspace,
- Regularizer $\Omega(\ldots)$ minimized when $f = f_s$.
- If $\Omega$ is strictly non-decreasing, then $\|f_{\perp}\|_{\mathcal{H}} = 0$ is required at the minimum.
Kernel ridge regression: proof

We begin knowing \( f \) is a linear combination of feature space mappings of points (representer theorem)

\[
f = \sum_{i=1}^{n} \alpha_i \phi(x_i).
\]

Then

\[
\sum_{i=1}^{n} (y_i - \langle f, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \| f \|^2_{\mathcal{H}} = \| y - K\alpha \|^2 + \lambda \alpha^\top K\alpha
\]

Differentiating wrt \( \alpha \) and setting this to zero, we get

\[
\alpha^* = (K + \lambda I_n)^{-1} y.
\]
Reminder: smoothness

What does $\|a\|_H$ have to do with smoothing?

**Example 1:** The Fourier series representation on torus $\mathbb{T}$:

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

and

$$\langle f, g \rangle_H = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l \hat{g}_l}{\hat{k}_l}.$$ 

Thus,

$$\|f\|_H^2 = \langle f, f \rangle_H = \sum_{l=-\infty}^{\infty} \frac{\hat{f}_l^2}{\hat{k}_l}.$$
Reminder: smoothness

What does $\|a\|_H$ have to do with smoothing?

**Example 2:** The Gaussian kernel on $\mathbb{R}$. Recall

$$f(x) = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i} e_i(x), \quad \|f\|_H^2 = \sum_{i=1}^{\infty} a_i^2.$$
Given the objective

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

How do we choose

- The regularization parameter $\lambda$?
- The kernel parameter: for Gaussian kernel, $\sigma$ in

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

- $\lambda = 0.1, \sigma = 0.6$
- $\lambda = 0.1, \sigma = 2$
- $\lambda = 0.1, \sigma = 0.1$