Foundations of Reproducing Kernel Hilbert Spaces II

Advanced Topics in Machine Learning

D. Sejdinovic, A. Gretton

Gatsby Unit

slides and notes are available at www.gatsby.ucl.ac.uk/~dino/teaching

March 17, 2014

The story so far

- **Hilbert space**:

The story so far

- **Hilbert space**: a complete space with an inner product

The story so far

- **Hilbert space**: a complete space with an inner product
    - *Riesz Theorem*:

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**:

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous
  - thus, evaluation is representable by an inner product

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous
  - thus, evaluation is representable by an inner product with *some element*

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous
  - thus, evaluation is representable by an inner product with *some element*
  - define $k(\cdot, x)$ as that representer of evaluation

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous
  - thus, evaluation is representable by an inner product with *some element*
  - define $k(\cdot, x)$ as that representer of evaluation: reproducing kernel

The story so far

- **Hilbert space**: a complete space with an inner product
  - *Riesz Theorem*: **all** linear & continuous functionals are representable by inner products
- **RKHS**: a Hilbert space of functions for which evaluation is continuous
  - thus, evaluation is representable by an inner product with *some element*
  - define $k(\cdot, x)$ as that representer of evaluation: reproducing kernel
- kernel as an inner product between features: $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

Overview

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

RKHS

Definition (Reproducing kernel Hilbert space)

Let $\mathcal{X}$ be a non-empty set. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if evaluation functionals $\delta_x : f \mapsto f(x)$ are continuous $\forall x \in \mathcal{X}$.

RKHS

Definition (Reproducing kernel Hilbert space)

Let $\mathcal{X}$ be a non-empty set. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if evaluation functionals $\delta_x : f \mapsto f(x)$ are continuous $\forall x \in \mathcal{X}$.

If two functions $f, g \in \mathcal{H}$ are close in the norm of $\mathcal{H}$, then $f(x)$ and $g(x)$ are close for all $x \in \mathcal{X}$

Outline

Will discuss three distinct concepts:

- reproducing kernel
- inner product between features
- positive definite function

Outline

Will discuss three distinct concepts:

- reproducing kernel
- inner product between features
- positive definite function

...and then show that they are all equivalent.

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X}; \nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Reproducing kernel

Definition (Reproducing kernel)

Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ 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

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

Reproducing kernel

Definition (Reproducing kernel)

Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ 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

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

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

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

Reproducing kernel of an RKHS

Theorem

If it exists, reproducing kernel is unique.

Theorem

$\mathcal{H}$ is a reproducing kernel Hilbert space if and only if it has a reproducing kernel.

Outline

Feature space inner product

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a kernel* on $\mathcal{X}$ if there exists a Hilbert space (not necessarilly an RKHS) $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$, such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.

Feature space inner product

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a kernel* on $\mathcal{X}$ if there exists a Hilbert space (not necessarilly an RKHS) $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$, such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.

- $\phi : \mathcal{X} \to \mathcal{F}$ is called a **feature map**,
- $\mathcal{F}$ is called a **feature space**.

Feature space inner product

Definition (Kernel)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *a kernel* on $\mathcal{X}$ if there exists a Hilbert space (not necessarilly an RKHS) $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$, such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.

- $\phi : \mathcal{X} \to \mathcal{F}$ is called a **feature map**,
- $\mathcal{F}$ is called a **feature space**.

Fact

Every ***reproducing kernel*** *is a* ***kernel*** *(every RKHS is a valid feature space).*

Non-uniqueness of feature representation

Example

Consider $\mathcal{X} = \mathbb{R}^2$, and $k(x,y) = \langle x, y\rangle^2$

$$\begin{aligned} k(x,y) &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1x_2y_1y_2 \\ &= \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}x_1x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}y_1y_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1^2 & x_2^2 & x_1x_2 & x_1x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ y_1y_2 \\ y_1y_2 \end{bmatrix}. \end{aligned}$$

so we can use the feature maps $\phi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$ *or* $\tilde{\phi}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1x_2 & x_1x_2 \end{bmatrix}$, with feature spaces $\mathcal{H} = \mathbb{R}^3$ *or* $\tilde{\mathcal{H}} = \mathbb{R}^4$.

Non-uniqueness of feature representation

Example

Consider $\mathcal{X} = \mathbb{R}^2$, and $k(x, y) = \langle x, y \rangle^2$

$$
\begin{aligned}
k(x, y) &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 x_2 y_1 y_2 \\
&= \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2} x_1 x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2} y_1 y_2 \end{bmatrix} \\
&= \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 & x_1 x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ y_1 y_2 \\ y_1 y_2 \end{bmatrix}.
\end{aligned}
$$

so we can use the feature maps $\phi(x) = \left(x_1^2, x_2^2, \sqrt{2} x_1 x_2\right)$ *or* $\tilde{\phi}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 & x_1 x_2 \end{bmatrix}$, with feature spaces $\mathcal{H} = \mathbb{R}^3$ *or* $\tilde{\mathcal{H}} = \mathbb{R}^4$.

Not RKHS!

Non-uniqueness of feature representation

Example

Consider $\mathcal{X} = \mathbb{R}^2$, and $k(x, y) = \langle x, y \rangle^2$

$$\begin{aligned} k(x, y) &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 \\ &= \begin{bmatrix} x_1^2 & x_2^2 & \sqrt{2}x_1 x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}y_1 y_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 & x_1 x_2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ y_2^2 \\ y_1 y_2 \\ y_1 y_2 \end{bmatrix}. \end{aligned}$$

so we can use the feature maps $\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$ *or* $\tilde{\phi}(x) = \begin{bmatrix} x_1^2 & x_2^2 & x_1 x_2 & x_1 x_2 \end{bmatrix}$, with feature spaces $\mathcal{H} = \mathbb{R}^3$ *or* $\tilde{\mathcal{H}} = \mathbb{R}^4$.

Not RKHS!

Evaluation is not defined on $\mathbb{R}^3$ or $\mathbb{R}^4$.

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Positive definite functions

Definition (Positive definite functions)

A symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* if $\forall n \geq 1,\ \forall (a_1, \ldots a_n) \in \mathbb{R}^n,\ \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) = \mathbf{a}^\top \mathbf{H} \mathbf{a} \geq 0.$$

Positive definite functions

Definition (Positive definite functions)

A symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* if $\forall n \geq 1$, $\forall (a_1, \ldots a_n) \in \mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) = \mathbf{a}^\top \mathbf{H} \mathbf{a} \geq 0.$$

h is *strictly* positive definite if for mutually distinct x_i, the equality holds only when all the a_i are zero.

Kernels are positive definite

Fact

Every kernel is a positive definite function.

Kernels are positive definite

Fact

Every kernel is a positive definite function.

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

So far

reproducing kernel $\implies$ kernel $\implies$ positive definite

So far

$$\text{reproducing kernel} \implies \text{kernel} \implies \text{positive definite}$$

Is every positive definite function a reproducing kernel for some RKHS?

So far

reproducing kernel $\Longleftrightarrow$ kernel $\Longleftrightarrow$ positive definite

Yes (Moore-Aronszajn)!

Outline

Moore-Aronszajn Theorem

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.*

Moore-Aronszajn Theorem: pre-RKHS

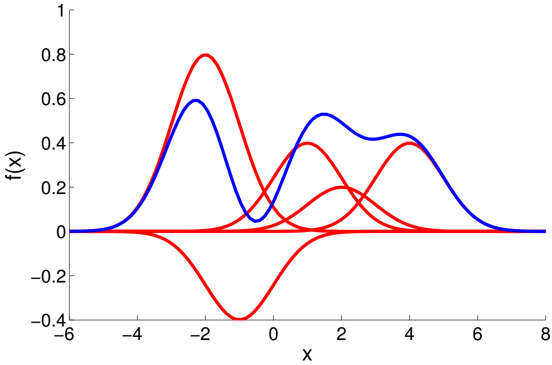
Starting with a positive def. k, construct a **pre-RKHS** (an inner product space) $\mathcal{H}_0 \subset \mathbb{R}^{\mathcal{X}}$ with properties:

1. The evaluation functionals δ_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 $\mathcal{H}_0 = span\{k(\cdot,x)\,|\,x \in \mathcal{X}\}$ will be taken to be the set of functions:

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



Moore-Aronszajn Theorem: Steps

Theorem (Moore-Aronszajn - Step A)

Space $\mathcal{H}_0 = span\{k(\cdot, x) \,|\, x \in \mathcal{X}\}$, *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.*

Moore-Aronszajn Theorem: Steps

Theorem (Moore-Aronszajn - Step A)

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

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

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 δ_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 $f \in \mathbb{R}^{\mathcal{X}}$ for which there exists an $\mathcal{H}_0$-Cauchy sequence $\{f_n\}$ converging **pointwise** to f. Clearly, $\mathcal{H}_0 \subseteq \mathcal{H}$.

Moore-Aronszajn Theorem- Step B

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. 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?

Moore-Aronszajn Theorem- Step B

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. 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?

Moore-Aronszajn Theorem- Step B

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. 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 $\mathcal{H}$?

Moore-Aronszajn Theorem- Step B

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. 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)?

Moore-Aronszajn Theorem- Step B

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. 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) $\implies \mathcal{H}$ is RKHS!

Summary

reproducing kernel $\iff$ kernel $\iff$ positive definite

Summary

$$\text{reproducing kernel} \iff \text{kernel} \iff \text{positive definite}$$

$$\text{set of all pd functions: } \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$$
$$\overset{1-1}{\longleftrightarrow}$$
$$\text{set of all RKHSs: } \mathit{Hilb}(\mathbb{R}^{\mathcal{X}})$$

Non-uniqueness of feature representation

- There are (infinitely) many feature space representations (and we can even work in one or more of them, if it's convenient!)

$$\langle \phi(x), \phi(y) \rangle_{\mathbb{R}^3} = a y_1^2 + b y_2^2 + c\sqrt{2} y_1 y_2 \quad = \quad k_x(y)$$

$$\phi(x) = \begin{bmatrix} a = & x_1^2 & b = x_2^2 & c = \sqrt{2} x_1 x_2 \end{bmatrix}$$

Non-uniqueness of feature representation

- There are (infinitely) many feature space representations (and we can even work in one or more of them, if it's convenient!)

$$\langle \phi(x), \phi(y) \rangle_{\mathbb{R}^3} = a y_1^2 + b y_2^2 + c\sqrt{2} y_1 y_2 \quad = \quad k_x(y)$$

$$\phi(x) = \begin{bmatrix} a = x_1^2 & b = x_2^2 & c = \sqrt{2} x_1 x_2 \end{bmatrix}$$

$$\left\langle \tilde{\phi}(x), \tilde{\phi}(y) \right\rangle_{\mathbb{R}^4} = \tilde{a} y_1^2 + \tilde{b} y_2^2 + \tilde{c} y_1 y_2 + \tilde{d} y_1 y_2 \quad = \quad k_x(y)$$

$$\tilde{\phi}(x) = \begin{bmatrix} \tilde{a} = x_1^2 & \tilde{b} = x_2^2 & \tilde{c} = x_1 x_2 & \tilde{d} = x_1 x_2 \end{bmatrix}$$

Non-uniqueness of feature representation

- There are (infinitely) many feature space representations (and we can even work in one or more of them, if it's convenient!)

$$\langle \phi(x), \phi(y) \rangle_{\mathbb{R}^3} = ay_1^2 + by_2^2 + c\sqrt{2}y_1y_2 \quad = \quad k_x(y)$$

$$\phi(x) = \begin{bmatrix} a = & x_1^2 & b = x_2^2 & c = \sqrt{2}x_1x_2 \end{bmatrix}$$

$$\left\langle \tilde{\phi}(x), \tilde{\phi}(y) \right\rangle_{\mathbb{R}^4} = \tilde{a}y_1^2 + \tilde{b}y_2^2 + \tilde{c}y_1y_2 + \tilde{d}y_1y_2 \quad = \quad k_x(y)$$

$$\tilde{\phi}(x) = \begin{bmatrix} \tilde{a} = x_1^2 & \tilde{b} = x_2^2 & \tilde{c} = x_1x_2 & \tilde{d} = x_1x_2 \end{bmatrix}$$

- Different feature maps give *coefficients* of canonical feature map $k(\cdot, x)$ in terms of (different) simpler functions.

Non-uniqueness of feature representation

- There are (infinitely) many feature space representations (and we can even work in one or more of them, if it's convenient!)

$$\langle \phi(x), \phi(y) \rangle_{\mathbb{R}^3} = a y_1^2 + b y_2^2 + c\sqrt{2} y_1 y_2 \quad = \quad k_x(y)$$

$$\phi(x) = \begin{bmatrix} a = & x_1^2 & b = x_2^2 & c = \sqrt{2} x_1 x_2 \end{bmatrix}$$

$$\left\langle \tilde{\phi}(x), \tilde{\phi}(y) \right\rangle_{\mathbb{R}^4} = \tilde{a} y_1^2 + \tilde{b} y_2^2 + \tilde{c} y_1 y_2 + \tilde{d} y_1 y_2 \quad = \quad k_x(y)$$

$$\tilde{\phi}(x) = \begin{bmatrix} \tilde{a} = x_1^2 & \tilde{b} = x_2^2 & \tilde{c} = x_1 x_2 & \tilde{d} = x_1 x_2 \end{bmatrix}$$

- Different feature maps give *coefficients* of canonical feature map $k(\cdot, x)$ in terms of (different) simpler functions.
- RKHS of k remains **unique**, regardless of the representation.

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X}; \nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Assumptions

- So far, no assumptions on:
  - $\mathcal{X}$ (apart from it being a non-empty set)
  - nor on k (apart from it being a positive definite function)

Assumptions

- So far, no assumptions on:
  - $\mathcal{X}$ (apart from it being a non-empty set)
  - nor on k (apart from it being a positive definite function)
- Now, assume that:

Assumptions

- So far, no assumptions on:
  - $\mathcal{X}$ (apart from it being a non-empty set)
  - nor on k (apart from it being a positive definite function)
- Now, assume that:
  - $\mathcal{X}$ is a compact metric space
    - such as $[a, b]^d$, **key**: *every continuous function on $\mathcal{X}$ is bounded and uniformly continuous*

Assumptions

- So far, no assumptions on:
  - $\mathcal{X}$ (apart from it being a non-empty set)
  - nor on k (apart from it being a positive definite function)
- Now, assume that:
  - $\mathcal{X}$ is a compact metric space
    - such as $[a, b]^d$, **key**: *every continuous function on $\mathcal{X}$ is bounded and uniformly continuous*
  - $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous positive definite function

Integral operator of a kernel

Definition (Integral operator)

Let ν be a finite Borel measure on $\mathcal{X}$. For the linear map

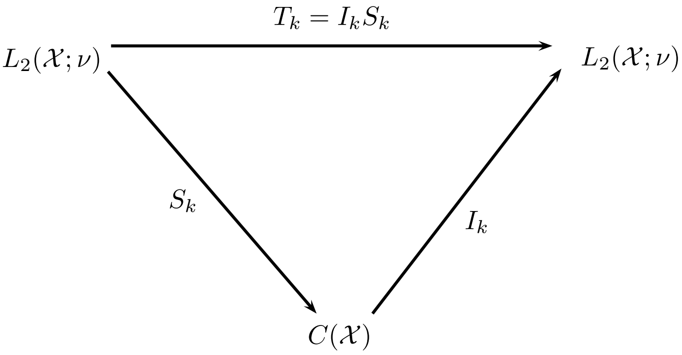
$$\begin{aligned} S_k : L_2(\mathcal{X};\nu) &\rightarrow \mathcal{C}(\mathcal{X}), \\ \left(S_k \tilde{f}\right)(x) &= \int k(x,y)f(y)d\nu(y), \quad f \in \tilde{f} \in L_2(\mathcal{X};\nu), \end{aligned}$$

its composition $T_k = I_k \circ S_k$ with the inclusion $I_k : \mathcal{C}(\mathcal{X}) \hookrightarrow L_2(\mathcal{X};\nu)$ is said to be the *integral operator* of k.

Proof that $S_k\tilde{f}$ is continuous

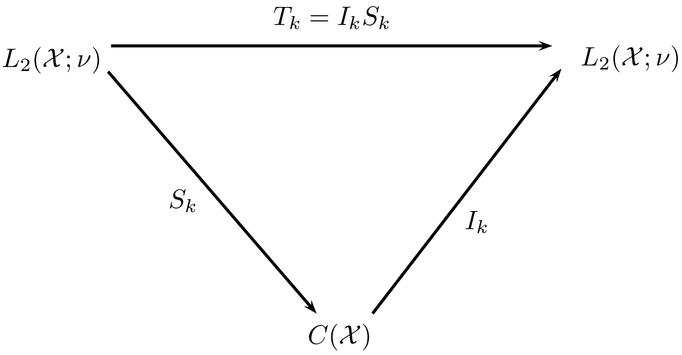
$$\begin{aligned}
\left|\left(S_k\tilde{f}\right)(x)-\left(S_k\tilde{f}\right)(x')\right| &= \left|\int \left(k(x,y)-k(x',y)\right) f(y)d\nu(y)\right| \\
&= \left|\left\langle I_k\left(k_x-k_{x'}\right),\tilde{f}\right\rangle_{L^2}\right| \\
&\leq \left\|I_k\left(k_x-k_{x'}\right)\right\|_{L^2}\left\|\tilde{f}\right\|_{L^2} \\
&= \left\|\tilde{f}\right\|_{L^2}\sqrt{\int \left(k(x,y)-k(x',y)\right)^2 d\nu(y)} \\
&\leq \nu(\mathcal{X})\left\|\tilde{f}\right\|_{L^2}\max_y\left|k(x,y)-k(x',y)\right|
\end{aligned}$$

Integral operator of a kernel (2)



$$T_k \;:\; L_2(\mathcal{X};\nu) \;\to\; L_2(\mathcal{X};\nu)$$

Integral operator of a kernel (2)



$$T_k \; : \; L_2(\mathcal{X};\nu) \;\; \to \;\; L_2(\mathcal{X};\nu)$$

$T_k \neq S_k$: $\left(S_k \tilde{f}\right)(x)$ is defined, while $\left(T_k \tilde{f}\right)(x)$ is **not**!

Properties of integral operators

- k symmetric $\implies$ T_k **self-adjoint**: $\langle f, T_k g \rangle = \langle T_k f, g \rangle$

Properties of integral operators

- k symmetric $\implies$ T_k **self-adjoint**: $\langle f, T_k g\rangle = \langle T_k f, g\rangle$
- k positive definite $\implies$ T_k **positive**: $\langle f, T_k f\rangle \geq 0$

Properties of integral operators

- k symmetric $\implies$ T_k **self-adjoint**: $\langle f, T_k g \rangle = \langle T_k f, g \rangle$
- k positive definite $\implies$ T_k **positive**: $\langle f, T_k f \rangle \geq 0$
- k continuous $\implies$ T_k **compact**: if $\{f_n\}$ is bounded, then $\{T_k f_n\}$ has a convergent subsequence

Properties of integral operators

- k symmetric $\implies$ T_k **self-adjoint**: $\langle f, T_k g\rangle = \langle T_k f, g\rangle$
- k positive definite $\implies$ T_k **positive**: $\langle f, T_k f\rangle \geq 0$
- k continuous $\implies$ T_k **compact**: if $\{f_n\}$ is bounded, then $\{T_k f_n\}$ has a convergent subsequence

Theorem (Spectral theorem)

Let $\mathcal{F}$ be a Hilbert space,and $T : \mathcal{F} \to \mathcal{F}$ a compact, self-adjoint operator. There is an at most countable ONS $\{u_j\}_{j\in J}$ of $\mathcal{F}$ and $\{\lambda_j\}_{j\in J}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ converging to zero such that

$$Tf = \sum_{j\in J} \lambda_j \langle f, u_j\rangle_{\mathcal{F}} u_j, \qquad f \in \mathcal{F}.$$

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X}; \nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$, so there exist ONS $\{\tilde{e}_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$, so there exist ONS $\{\tilde{e}_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$ (strictly positive eigenvalues; J at most countable).

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$, so there exist ONS $\{\tilde{e}_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$ (strictly positive eigenvalues; J at most countable).
- $\tilde{e}_j$ is an equivalence class in the ONS of $L_2(\mathcal{X}; \nu)$

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$, so there exist ONS $\{\tilde{e}_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$ (strictly positive eigenvalues; J at most countable).
- $\tilde{e}_j$ is an equivalence class in the ONS of $L_2(\mathcal{X}; \nu)$
- $e_j = \lambda_j^{-1} S_k \tilde{e}_j \in \mathcal{C}(\mathcal{X})$ is a continuous function in the class $\tilde{e}_j$: $I_k e_j = \tilde{e}_j$.

Mercer's theorem

- $\mathcal{X}$ a compact metric space; $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel.
- A finite measure ν on $\mathcal{X}$ with $supp \nu = \mathcal{X}$.
- Integral operator T_k is then compact, positive and self-adjoint on $L_2(\mathcal{X}; \nu)$, so there exist ONS $\{\tilde{e}_j\}_{j \in J}$ and $\{\lambda_j\}_{j \in J}$ (strictly positive eigenvalues; J at most countable).
- $\tilde{e}_j$ is an equivalence class in the ONS of $L_2(\mathcal{X}; \nu)$
- $e_j = \lambda_j^{-1} S_k \tilde{e}_j \in \mathcal{C}(\mathcal{X})$ is a continuous function in the class $\tilde{e}_j$: $I_k e_j = \tilde{e}_j$.

Theorem (Mercer's theorem)

$\forall x, y \in \mathcal{X}$ with convergence uniform on $\mathcal{X} \times \mathcal{X}$:

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

Mercer's theorem (2)

$$
\begin{aligned}
k(x,y) &= \sum_{j\in J} \lambda_j e_j(x) e_j(y) \\
&= \left\langle \left\{ \sqrt{\lambda_j} e_j(x) \right\}, \left\{ \sqrt{\lambda_j} e_j(y) \right\} \right\rangle_{\ell^2(J)}
\end{aligned}
$$

Mercer's theorem (2)

$$
\begin{aligned}
k(x,y) &= \sum_{j\in J} \lambda_j e_j(x) e_j(y) \\
&= \left\langle \left\{ \sqrt{\lambda_j} e_j(x) \right\}, \left\{ \sqrt{\lambda_j} e_j(y) \right\} \right\rangle_{\ell^2(J)}
\end{aligned}
$$

Another (Mercer) feature map:

$$
\begin{aligned}
\phi : \mathcal{X} &\to \ell^2(J) \\
\phi : x &\mapsto \left\{ \sqrt{\lambda_j} e_j(x) \right\}_{j\in J}
\end{aligned}
$$

Mercer's theorem (2)

$$
\begin{aligned}
k(x,y) &= \sum_{j\in J} \lambda_j e_j(x) e_j(y) \\
&= \left\langle \left\{ \sqrt{\lambda_j} e_j(x) \right\}, \left\{ \sqrt{\lambda_j} e_j(y) \right\} \right\rangle_{\ell^2(J)}
\end{aligned}
$$

Another (Mercer) feature map:

$$
\begin{aligned}
\phi : \mathcal{X} &\to \ell^2(J) \\
\phi : x &\mapsto \left\{ \sqrt{\lambda_j} e_j(x) \right\}_{j\in J}
\end{aligned}
$$

$$
\sum_{j\in J} \left(\sqrt{\lambda_j} e_j(x) \right)^2 = k(x,x) < \infty
$$

Mercer's theorem (3)

- Sum $\sum_{j\in J} a_j e_j(x)$ converges absolutely $\forall x \in \mathcal{X}$ whenever sequence $\left\{a_j/\sqrt{\lambda_j}\right\} \in \ell^2(J)$:

$$\begin{aligned}\sum_{j\in J} \left|a_j e_j(x)\right| &\leq \left[\sum_{j\in J} \left|a_j/\sqrt{\lambda_j}\right|^2\right]^{1/2} \cdot \left[\sum_{j\in J} \left|\sqrt{\lambda_j} e_j(x)\right|^2\right]^{1/2} \\ &= \left\|\left\{a_j/\sqrt{\lambda_j}\right\}\right\|_{\ell^2(J)} \sqrt{k(x,x)}.\end{aligned}$$

Mercer's theorem (3)

- Sum $\sum_{j \in J} a_j e_j(x)$ converges absolutely $\forall x \in \mathcal{X}$ whenever sequence $\left\{a_j/\sqrt{\lambda_j}\right\} \in \ell^2(J)$:

$$\begin{aligned} \sum_{j \in J} |a_j e_j(x)| &\leq \left[\sum_{j \in J} \left|a_j/\sqrt{\lambda_j}\right|^2\right]^{1/2} \cdot \left[\sum_{j \in J} \left|\sqrt{\lambda_j} e_j(x)\right|^2\right]^{1/2} \\ &= \left\|\left\{a_j/\sqrt{\lambda_j}\right\}\right\|_{\ell^2(J)} \sqrt{k(x,x)}. \end{aligned}$$

$\sum_{j \in J} a_j e_j$ is a well defined function on $\mathcal{X}$

Mercer representation of RKHS

Theorem

Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel. Define:

$$\mathcal{H} = \left\{ f = \sum_{j \in J} a_j e_j \; : \; \left\{ a_j / \sqrt{\lambda_j} \right\} \in \ell^2(J) \right\},$$

with inner product:

$$\left\langle \sum_{j \in J} a_j e_j, \sum_{j \in J} b_j e_j \right\rangle_{\mathcal{H}} = \sum_{j \in J} \frac{a_j b_j}{\lambda_j}.$$

Then $\mathcal{H}$ is the RKHS of k.

Mercer representation of RKHS

Theorem

Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel. Define:

$$\mathcal{H} = \left\{ f = \sum_{j \in J} a_j e_j \; : \; \left\{ a_j / \sqrt{\lambda_j} \right\} \in \ell^2(J) \right\},$$

with inner product:

$$\left\langle \sum_{j \in J} a_j e_j, \sum_{j \in J} b_j e_j \right\rangle_{\mathcal{H}} = \sum_{j \in J} \frac{a_j b_j}{\lambda_j}.$$

Then $\mathcal{H}$ is the RKHS of k.

RKHS is unique, so does not depend on ν !

Proof

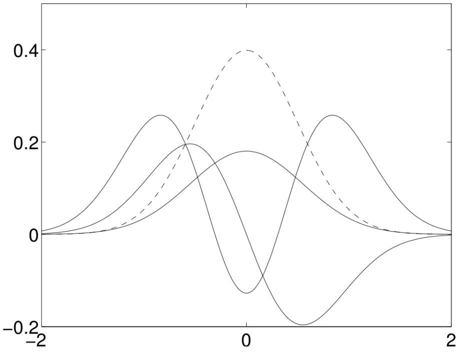
1. $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product: if $f=\sum_{j\in J}a_je_j$ then$\langle f,f\rangle_{\mathcal{H}}=\sum_{j\in J}\frac{a_j^2}{\lambda_j}>0$ if some $a_j>0$
2. Let $\{f_n\}$ be Cauchy, $f_n=\sum_{j\in J}a_j^{(n)}e_j$. Then $\|f_n-f_m\|_{\mathcal{H}}^2=\sum_{j\in J}\frac{\left(a_j^{(n)}-a_j^{(m)}\right)^2}{\lambda_j}=\left\|\left\{a_j^{(n)}/\sqrt{\lambda_j}\right\}-\left\{a_j^{(m)}/\sqrt{\lambda_j}\right\}\right\|_{\ell^2}^2<\epsilon$, so must have a limit because ℓ^2 is a Hilbert space.
3. $k(\cdot,x)=\sum_{j\in J}[\lambda_je_j(x)]\,e_j\in\mathcal{H}$ since $\sum_{j\in J}\left(\frac{\lambda_je_j(x)}{\sqrt{\lambda_j}}\right)^2=k(x,x)<\infty$
4. $\langle f,k(\cdot,x)\rangle_{\mathcal{H}}=\left\langle\sum_{j\in J}a_je_j,\sum_{j\in J}[\lambda_je_j(x)]\,e_j\right\rangle_{\mathcal{H}}=\sum_{j\in J}\frac{a_j\lambda_je_j(x)}{\lambda_j}=\sum_{j\in J}a_je_j(x)=f(x)$.

Smoothness interpretation

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

$$\lambda_j \propto b^j \qquad b < 1$$
$$e_j(x) \propto \exp(-(c-a)x^2)H_j(x\sqrt{2c}),$$

a, b, c are functions of σ, and H_j is jth order Hermite polynomial.



NOTE that $\|f\|_{\mathcal{H}_k} < \infty$ is a "smoothness" constraint: λ_j decay as e_j become "rougher" and

$$\|f\|^2_{\mathcal{H}_k} = \sum_{j \in J} \frac{a_j^2}{\lambda_j}$$

(Figure from Rasmussen and Williams)

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

Assume $\{\tilde{e}_j\}_{j\in J}$ is ONB of $L_2(\mathcal{X};\nu)$, and write $\hat{f}(j) = \langle f, \tilde{e}_j \rangle_{L_2}$

$$T_k f = \sum_{j\in J} \lambda_j \hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu)$$

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

Assume $\{\tilde{e}_j\}_{j\in J}$ is ONB of $L_2(\mathcal{X};\nu)$, and write $\hat{f}(j) = \langle f, \tilde{e}_j \rangle_{L_2}$

$$T_k f = \sum_{j\in J} \lambda_j \hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu)$$

$$T_k^{1/2} f = \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu)$$

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

Assume $\{\tilde{e}_j\}_{j\in J}$ is ONB of $L_2(\mathcal{X};\nu)$, and write $\hat{f}(j) = \langle f, \tilde{e}_j\rangle_{L_2}$

$$T_k f = \sum_{j\in J} \lambda_j \hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu)$$

$$T_k^{1/2} f = \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)\tilde{e}_j, \qquad f \in L_2(\mathcal{X};\nu)$$

$$\mathcal{H}_k = \left\{ f = \sum_{j\in J} a_j e_j \ : \ \left\{ a_j/\sqrt{\lambda_j} \right\} \in \ell^2(J) \right\}$$

$$\sum_{j\in J} \left|\hat{f}(j)\right|^2 = \|f\|_2^2 < \infty \Rightarrow \left\{\hat{f}(j)\right\} \in \ell^2(J) \ \Rightarrow \ \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j) e_j \in \mathcal{H}_k$$

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

$$\langle f, g\rangle_{L_2} = \left\langle \left\{\hat{f}(j)\right\}, \{\hat{g}(j)\}\right\rangle_{\ell^2(J)} = \left\langle \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j, \sum_{j\in J} \sqrt{\lambda_j}\hat{g}(j)e_j \right\rangle_{\mathcal{H}_k}$$

$\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

$$\langle f, g\rangle_{L_2} = \left\langle \left\{\hat{f}(j)\right\}, \{\hat{g}(j)\} \right\rangle_{\ell^2(J)} = \left\langle \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j, \sum_{j\in J} \sqrt{\lambda_j}\hat{g}(j)e_j \right\rangle_{\mathcal{H}_k}$$

$T_k^{1/2}$ induces an isometric isomorphism between $span\{\tilde{e}_j \ : j \in J\} \subseteq L_2(\mathcal{X};\nu)$ and $\mathcal{H}_k$ (and both are isometrically isomorphic to $\ell^2(J)$).

Canonical feature map

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

Canonical feature map

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

$$k(\cdot, x) = \sum_{j\in J} \sqrt{\lambda_j} \left(\sqrt{\lambda_j} e_j(x)\right) e_j$$

Canonical feature map

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j)e_j \in \mathcal{H}_k$$

$$k(\cdot,x) = \sum_{j\in J} \sqrt{\lambda_j}\left(\sqrt{\lambda_j}e_j(x)\right)e_j$$

$$\mathcal{H}_k \ni k(\cdot,x) \leftarrow x \rightarrow \left\{\sqrt{\lambda_j}e_j(x)\right\} \in \ell^2(J)$$

Canonical feature map

$$f \in L_2(\mathcal{X};\nu) \overset{1-1}{\longleftrightarrow} \left\{\hat{f}(j)\right\} \in \ell^2(J) \quad \overset{1-1}{\longleftrightarrow} \quad \sum_{j\in J} \sqrt{\lambda_j}\hat{f}(j) e_j \in \mathcal{H}_k$$

$$k(\cdot,x) = \sum_{j\in J} \sqrt{\lambda_j}\left(\sqrt{\lambda_j} e_j(x)\right) e_j$$

$$\mathcal{H}_k \ni k(\cdot,x) \leftarrow x \rightarrow \left\{\sqrt{\lambda_j} e_j(x)\right\} \in \ell^2(J)$$

Mercer feature map gives Fourier coefficients of the canonical feature map.

Outline

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X}; \nu)$
4. Operations with kernels
   - **Sum and product**
   - Constructing new kernels

Operations with kernels

Fact (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.

Operations with kernels

Fact (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.

- A difference of kernels is not necessarily a kernel! This is because we cannot have $k_1(x,x) - k_2(x,x) = \langle \phi(x), \phi(x) \rangle_{\mathcal{H}} < 0$.
- This gives the set of all kernels the geometry of a *closed convex cone.*

Operations with kernels

Fact (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.

- A difference of kernels is not necessarily a kernel! This is because we cannot have $k_1(x,x) - k_2(x,x) = \langle \phi(x), \phi(x) \rangle_{\mathcal{H}} < 0$.
- This gives the set of all kernels the geometry of a *closed convex cone.*

$$\mathcal{H}_{k_1+k_2} = \mathcal{H}_{k_1} + \mathcal{H}_{k_2} = \{f_1 + f_2 \,:\, f_1 \in \mathcal{H}_{k_1}, f_2 \in \mathcal{H}_{k_2}\}$$

Operations with kernels (2)

Fact (Product of kernels)

If k_1 and k_2 are kernels on $\mathcal{X}$ and $\mathcal{Y}$, then $k = k_1 \otimes k_2$, given by:

$$k\left((x, y), (x', y')\right) := k_1(x, x')k_2(y, y')$$

is a kernel on $\mathcal{X} \times \mathcal{Y}$. If $\mathcal{X} = \mathcal{Y}$, then $k = k_1 \cdot k_2$, given by:

$$k\left(x, x'\right) := k_1(x, x')k_2(x, x')$$

is a kernel on $\mathcal{X}$.

Operations with kernels (2)

Fact (Product of kernels)

If k_1 and k_2 are kernels on $\mathcal{X}$ and $\mathcal{Y}$, then $k = k_1 \otimes k_2$, given by:

$$k\left((x,y),(x',y')\right) \quad := \quad k_1(x,x')k_2(y,y')$$

is a kernel on $\mathcal{X} \times \mathcal{Y}$. If $\mathcal{X} = \mathcal{Y}$, then $k = k_1 \cdot k_2$, given by:

$$k\left(x,x'\right) \quad := \quad k_1(x,x')k_2(x,x')$$

is a kernel on $\mathcal{X}$.

$$\mathcal{H}_{k_1 \otimes k_2} \cong \mathcal{H}_{k_1} \otimes \mathcal{H}_{k_2}$$

Summary

$$\text{all kernels } \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}} \overset{1-1}{\longleftrightarrow} \text{all function spaces with continuous evaluation } \mathit{Hilb}(\mathbb{R}^{\mathcal{X}})$$

Summary

$$\text{all kernels } \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$$

$$\overset{1-1}{\longleftrightarrow}$$

$$\text{all function spaces with continuous evaluation } \mathit{Hilb}(\mathbb{R}^{\mathcal{X}})$$

bijection between $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$ and $\mathit{Hilb}(\mathbb{R}^{\mathcal{X}})$ preserves geometric structure

Outline

1. What is an RKHS?
   - Reproducing kernel
   - Inner product between features
   - Positive definite function
2. Moore-Aronszajn Theorem
3. Mercer representation of RKHS
   - Integral operator
   - Mercer's theorem
   - Relation between $\mathcal{H}_k$ and $L_2(\mathcal{X};\nu)$
4. Operations with kernels
   - Sum and product
   - Constructing new kernels

Kernels on $\mathbb{R}^d$

New kernels from old:

- trivial (linear) kernel on $\mathbb{R}^d$ is $k(x, x') = \langle x, x' \rangle$

Kernels on $\mathbb{R}^d$

New kernels from old:

- trivial (linear) kernel on $\mathbb{R}^d$ is $k(x, x') = \langle x, x' \rangle$
- for any $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with $a_i \geq 0$
  $\implies k(x, x') = p(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$

Kernels on $\mathbb{R}^d$

New kernels from old:

- trivial (linear) kernel on $\mathbb{R}^d$ is $k(x, x') = \langle x, x' \rangle$
- for any $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with $a_i \geq 0$
  $\implies k(x, x') = p(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$
- **polynomial kernel:** $k(x, x') = (\langle x, x' \rangle + c)^m$, for $c \geq 0$

Kernels on $\mathbb{R}^d$

New kernels from old:

- trivial (linear) kernel on $\mathbb{R}^d$ is $k(x, x') = \langle x, x' \rangle$
- for any $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with $a_i \geq 0$
  $\implies k(x, x') = p(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$
- **polynomial kernel:** $k(x, x') = (\langle x, x' \rangle + c)^m$, for $c \geq 0$
- $f(t)$ has Taylor series with non-negative coefficients
  $\implies k(x, x') = f(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$

Kernels on $\mathbb{R}^d$

New kernels from old:

- trivial (linear) kernel on $\mathbb{R}^d$ is $k(x, x') = \langle x, x' \rangle$
- for any $p(t) = a_m t^m + \cdots + a_1 t + a_0$ with $a_i \geq 0$
  $\implies k(x, x') = p(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$
- **polynomial kernel:** $k(x, x') = (\langle x, x' \rangle + c)^m$, for $c \geq 0$
- $f(t)$ has Taylor series with non-negative coefficients
  $\implies k(x, x') = f(\langle x, x' \rangle)$ is a kernel on $\mathbb{R}^d$
- **exponential kernel:** $k(x, x') = \exp(\sigma \langle x, x' \rangle)$, for $\sigma > 0$

Kernels on $\mathbb{R}^d$

New kernels from old:

- **polynomial kernel:** $k(x, x') = (\langle x, x' \rangle + c)^m$, for $c \geq 0$
- **exponential kernel:** $k(x, x') = \exp(\sigma \langle x, x' \rangle)$, for $\sigma > 0$

Gaussian kernel

Let $\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(x) = \exp(-\sigma \|x\|^2)$. Then, $\tilde{k}$ is representable as an inner product in $\mathbb{R}$:

$$\tilde{k}(x, x') = \phi(x)\phi(x') = \exp(-\sigma \|x\|^2) \exp(-\sigma \|x'\|^2) \qquad \text{kernel!}$$

Gaussian kernel

Let $\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(x) = \exp(-\sigma \|x\|^2)$. Then, $\tilde{k}$ is representable as an inner product in $\mathbb{R}$:

$$\tilde{k}(x, x') = \phi(x)\phi(x') = \exp(-\sigma \|x\|^2)\exp(-\sigma \|x'\|^2) \qquad \text{kernel!}$$

$$\begin{aligned} k_{gauss}(x, x') &= \tilde{k}(x, x')k_{exp}(x, x') \\ &= \exp\left(-\sigma\left[\|x\|^2 + \|x'\|^2 - 2\langle x, x'\rangle\right]\right) \\ &= \exp\left(-\sigma \|x - x'\|^2\right) \qquad \text{kernel!} \end{aligned}$$

Moore-Aronszajn Theorem

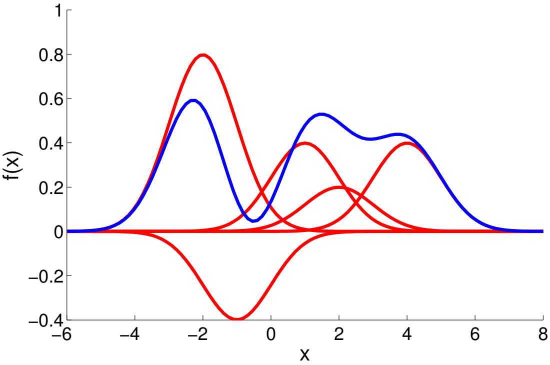
Starting with a positive def. k, construct a **pre-RKHS** (an inner product space of functions) $\mathcal{H}_0 \subset \mathbb{R}^{\mathcal{X}}$ with properties:

1. The evaluation functionals δ_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 (2)

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

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



Moore-Aronszajn Theorem (3)

Theorem (Moore-Aronszajn - Step I)

Space $\mathcal{H}_0 = span\{k(\cdot, x) \,|\, x \in \mathcal{X}\}$, *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.*

Moore-Aronszajn Theorem (3)

Theorem (Moore-Aronszajn - Step I)

Space $\mathcal{H}_0 = span\{k(\cdot, x) \,|\, x \in \mathcal{X}\}$*, 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 II)

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 (4)

Theorem (Moore-Aronszajn - Step I)

Space $\mathcal{H}_0 = span\left\{k(\cdot, x) \,|\, x \in \mathcal{X}\right\}$, *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.*

Moore-Aronszajn Theorem (4)

Theorem (Moore-Aronszajn - Step I)

Space $\mathcal{H}_0 = span\left\{k(\cdot,x) \,|\, x \in \mathcal{X}\right\}$, *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.*

1. The evaluation functionals δ_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 (5)

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. Clearly, $\mathcal{H}_0 \subseteq \mathcal{H}$.

Moore-Aronszajn Theorem (5)

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. 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?

Moore-Aronszajn Theorem (5)

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. 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?

Moore-Aronszajn Theorem (5)

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. 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 $\mathcal{H}$?

Moore-Aronszajn Theorem (5)

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. 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)?

Moore-Aronszajn Theorem (5)

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. 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) $\Longrightarrow \mathcal{H}$ is RKHS!

Summary

$$\text{reproducing kernel} \iff \text{kernel} \iff \text{positive definite}$$

Summary

$$\text{reproducing kernel} \iff \text{kernel} \iff \text{positive definite}$$

$$\text{all pd functions } \mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$$
$$\overset{1-1}{\longleftrightarrow}$$
$$\text{all function spaces with continuous evaluation } \mathit{Hilb}(\mathbb{R}^{\mathcal{X}})$$