STAT G8325 Gaussian Processes and Kernel Methods §08: Reproducing Kernel Hilbert Spaces

John P. Cunningham

Department of Statistics
Columbia University

Outline

Outline

Progress...

Week	Lectures	Content
7	Nov 4,9	Bayesian optimization and active learning
8	Nov 9, 11	Reproducing kernel Hilbert spaces
		• [Wah90, ch. 1] (intentionally light reading; work on projects)
9		Introduction to kernel methods

- HW3 due yesterday.
- HW4 due next Friday. Choose either:
  - complete introduction, background, literature review.
  - complete a code prototype, initial proof of concept.
- Who will be here Wednesday Nov 25?

Attribution

- The following sections introduce important concepts to gaussian processes and kernel methods more generally.
- We cover basic topics from functional analysis, and their applications.
- There are numerous reviews/introductions/texts.
- As such, the following draws heavily from:
  - Arthur Gretton [Gre13]
  - Dino Sejdinovic [SG12]
  - Christopher Heil [Hei06]
  - Sayan Mukherjee [Muk15]
  - Terry Tao [Tao09]
  - ... plus a few key textbooks [Kre89]; [TL58]; [SC08].
- These modern technical reports and lecture notes have clear examples and an appealing machine learning orientation.

Outline

Vector space

- We restrict our interest to a vector space $\mathcal{V}$ over the field of real numbers $\mathbb{R}$.
- As a reminder, a vector space is a set $\mathcal{V}$ with:
  - (using $f, g, h, 0 \in \mathcal{V}$ and $\alpha, \beta, 1 \in \mathbb{R}$)
  - additivity:

$$\begin{aligned} f + (g + h) &= (f + g) + h \\ f + g &= g + f \\ f + 0 &= f \\ f + -f &= 0 \end{aligned}$$

  - scalar multiplication:

$$\begin{aligned} \alpha(\beta f) &= (\alpha\beta) f \\ 1 f &= f \\ \alpha(f + g) &= \alpha f + \alpha g \\ (\alpha + \beta) f &= \alpha f + \beta f \end{aligned}$$

- Nothing unusual here.

Normed space

- A vector space $\mathcal{V}$ is a *normed space* if $\forall f \in \mathcal{V}$, there exists $\|f\| \in \mathbb{R}$ with:
  i. $\|f\| \geq 0$,
  ii. $\|f\| = 0 \Leftrightarrow f = 0$,
  iii. $\|\alpha f\| = |\alpha| \|f\| \quad \forall \alpha \in \mathbb{R}$,
  iv. $\|f + g\| \leq \|f\| + \|g\|$.
- If $\mathcal{V}$ is a normed space, with a sequence $\{f_n\}_{n\in\mathbb{N}}, f_n \in \mathcal{V}$:
  - We say $\{f_n\}_{n\in\mathbb{N}}$ *converges* to $f \in \mathcal{V}$ if:

$$\lim_{n\to\infty} \|f - f_n\| = 0 \qquad \Leftrightarrow \qquad \forall \epsilon > 0,\ \exists N \text{ such that } \forall n \geq N,\ \|f - f_n\| < \epsilon.$$

  - We say $\{f_n\}_{n\in\mathbb{N}}$ is *Cauchy* if:

$$\forall \epsilon > 0,\ \exists N \text{ such that } \forall n, m \geq N,\ \|f_m - f_n\| < \epsilon.$$

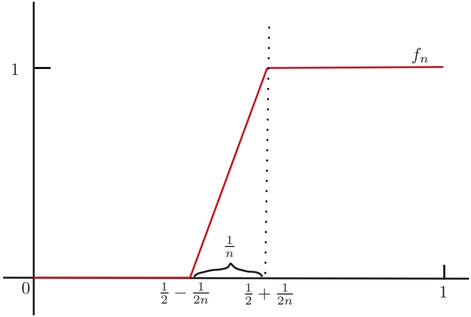
- Convergent sequences are Cauchy; Cauchy need not imply convergent:

$$\|f_m - f_n\| \leq \|f_m - f\| + \|f - f_n\|.$$

- This distinction is relevant...

Cauchy sequences need not be convergent

- Consider the normed space $\{\mathbb{Q}, |\cdot|\}$, and the sequence $1, 1.4, 1.41, \ldots$.
  - This sequence is Cauchy... for any $\epsilon > 0$, choose N such that $\epsilon > 10^{-N}$.
  - Then we have: $\forall n, m \geq N, \; \|f_m - f_n\| < \epsilon$.
  - This sequence is not convergent: the limit is $\sqrt{2} \notin \mathbb{Q}$.
- Take $C^{[0,1]}$, all continuous functions on $[0,1]$, with $\|f\| = \sqrt{\int_0^1 |f(x)|^2 dx}$.
- The sequence of functions below is again Cauchy, but with limit $f \notin C^{[0,1]}$.



Banach space

- A normed vector space for which all Cauchy sequences are convergent is called *complete*.
- A *Banach space* is a complete normed space; it contains the limits of all Cauchy sequences in that space.
- Some examples of Banach spaces (without proof):

$$
\begin{aligned}
L_p(\mathbb{R}) &= \left\{ f : \mathbb{R} \to \mathbb{R}, \int_{\mathbb{R}} |f(x)|^p dx < \infty \right\}, & \|f\|_p &= \left(\int |f(x)|^p dx \right)^{\frac{1}{p}}. \\
L_\infty(\mathbb{R}) &= \left\{ f : \mathbb{R} \to \mathbb{R}, \quad f \text{ essentially bounded} \right\}, & \|f\|_p &= \text{esssup}_{x \in \mathbb{R}} |f(x)|. \\
C_b(\mathbb{R}) &= \left\{ f \in L_\infty(\mathbb{R}), \quad f \text{ bounded and continuous} \right\}, & \|f\|_\infty &= \text{sup}_{x \in \mathbb{R}} |f(x)|. \\
C_0(\mathbb{R}) &= \left\{ f \in C_b(\mathbb{R}), \quad \lim_{|x| \to \infty} f(x) = 0 \right\}, & \|f\|_\infty &= \text{sup}_{x \in \mathbb{R}} |f(x)|.
\end{aligned}
$$

...recall esssup excludes points of zero measure

- Closed subspaces of Banach spaces are also Banach spaces. Examples:
  - $C_b(\mathbb{R})$ and $C_0(\mathbb{R})$ are closed subspaces of $L_\infty(\mathbb{R})$ with ℓ_∞ norm.

Hilbert space

- A vector spaced $\mathcal{V}$ is an *inner product space* if $\forall f, g \in \mathcal{V},\ \exists \langle f, g \rangle$ with:
  i. $\langle f, g \rangle = \overline{\langle g, f \rangle}$ (...which implies $\langle f, f \rangle \in \mathbb{R}$).
  ii. $\langle f, f \rangle \geq 0$,
  iii. $\langle f, f \rangle = 0 \quad \Rightarrow \quad f = 0$,
  iv. $\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle$.

- Some additional facts:
  - An *induced norm* is $\|f\| = \langle f, f \rangle^{\frac{1}{2}}$.
  - Thus all inner product spaces are normed spaces.
  - *Cauchy-Schwartz inequality*: $|\langle f, g \rangle| \leq \|f\| \|g\|$.
  - *Parallelogram rule*: $\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2$.
  - *Polarization identity*: $4 \langle f, g \rangle = \|f + g\|^2 - \|f - g\|^2$.

- A *Hilbert space* is a complete inner product space.

- A Hilbert space is a Banach space with norm induced by an inner product.

- Signpost: remember a kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$. Hilbert spaces will help us properly understand kernels.

Examples of Hilbert spaces

- Euclidean space:

$$\mathbb{R}^d, \text{ with } \langle f, g \rangle = \sum_{i=1}^{d} f_i g_i \;\forall f, g \in \mathbb{R}^d.$$

- $\ell_2(S)$, the set of square summable sequences of a countable index set S:

$$\{f_i\}_{i \in S}, \text{ such that } f_i \in \mathbb{R} \text{ and } \sum_{i \in S} |f_i|^2 < \infty, \text{ with } \langle \{f_i\}, \{g_i\} \rangle = \sum_{i \in S} f_i g_i.$$

- $L_2(\mathcal{X}, \mu)$, the set of all square integrable functions:

$$\begin{aligned} L_2(\mathcal{X}, \mu) &\triangleq \left\{ f : \mathcal{X} \to \mathbb{R} \text{ and measurable, with } \|f\|_2 = \left(\int_{\mathcal{X}} |f(x)|^2 d\mu \right)^{\frac{1}{2}} < \infty \right\}, \\ \text{with } \langle f, g \rangle &= \int_{\mathcal{X}} f(x) g(x) d\mu. \end{aligned}$$

- $L_2(\mathcal{X})$ typically means implied Lebesgue measure $\langle f, g \rangle = \int_{\mathcal{X}} f(x) g(x) dx$.

Separability

- Separability is a detail that is often skipped or assumed.
- We will revisit it later when considering rkhs, but for now we just define it and offer intuition.
- Consider a subspace $\mathcal{S}$ of a Banach space $\mathcal{V}$:
  - The *closure* $\bar{\mathcal{S}}$ is the union of $\mathcal{S}$ and all limit points (limits of sequences in $\mathcal{S}$).
  - $\mathcal{S}$ is *dense* in $\mathcal{V}$ if and only if $\bar{\mathcal{S}} = \mathcal{V}$.
  - Example: $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$.
  - A normed space $\mathcal{V}$ is *separable* if and only if $\exists$ a countable dense subset of $\mathcal{V}$.
- Separable Hilbert spaces have countable orthonormal bases.
  - This means that we can very (very!) loosely consider a Hilbert space to be intuitively like (possibly infinite dimensional) Euclidean space.
  - More rigorously, any separable infinite dimensional Hilbert space is isometrically isomorphic to $\ell_2(\mathbb{N})$ (i.e., square summable sequences).
  - We will sometimes assume separability.

Operators and basic definitions

- *Operator*: a map from one vector space to another.
- *Linear operator*: a map $L : \mathcal{V} \to \mathcal{H}$ obeying superposition and homogeneity:

$$\begin{aligned} L(f+g) &= Lf + Lg \qquad \forall f, g \in \mathcal{V} \\ L(\alpha f) &= \alpha L f \qquad \forall f \in \mathcal{V}, \alpha \in \mathbb{R} \end{aligned}$$

- *Continuous (at a point) operator*: at some point $f_0 \in \mathcal{V}$:

$$\forall \epsilon > 0, \quad \exists \delta(\epsilon, f_0) > 0 \text{ such that } \|f - f_0\|_{\mathcal{V}} < \delta(\epsilon, f_0) \Rightarrow \|Lf - Lf_0\|_{\mathcal{H}} < \epsilon.$$

- *Continuous operator*: an operator that is continuous at all points $f_0 \in \mathcal{V}$.
- *Uniformly continuous operator*: $\delta(\epsilon, f_0) = \delta(\epsilon)$, i.e. independent of f_0.
- *Lipschitz continuous operator*:

$$\exists K > 0 \text{ such that } \forall f_1, f_2 \in \mathcal{V}, \|Lf_1 - Lf_2\|_{\mathcal{H}} \leq K \|f_1 - f_2\|_{\mathcal{V}}.$$

- *Bounded operator*: an operator L is bounded if it has finite *operator norm*:

$$\|L\| = \sup_{f \in \mathcal{V}} \frac{\|Lf\|_{\mathcal{H}}}{\|f\|_{\mathcal{V}}} < \infty.$$

...L maps bounded subsets in $\mathcal{V}$ to bounded subsets in $\mathcal{H}$.

- Linear operator L: continuous a.a.p. $\Leftrightarrow$ continuous $\Leftrightarrow$ bounded.

Riesz representation theorem

- *Functional*: an operator that maps to $\mathbb{R}$, namely $L : \mathcal{V} \to \mathbb{R}$.
- *(Riesz representation theorem): in a Hilbert space $\mathcal{V}$, all continuous linear functionals L are inner products $\langle w, \cdot \rangle_{\mathcal{V}} : \mathcal{V} \to \mathbb{R}$, where $w \in \mathcal{V}$. In other words, $Lv = \langle w, v \rangle_{\mathcal{V}}$.*
  - If you are still thinking in Euclidean space, this is obvious.
  - More generally, it is not at all obvious.
  - Riesz representation theorem is **not** the representer theorem (coming later).
  - Riesz helps us define kernels using linear functionals in a Hilbert space.
- *Dual space*: all continuous linear functionals $\mathcal{V}' = \{\phi_w = \langle w, \cdot \rangle_{\mathcal{V}} : \mathcal{V} \to \mathbb{R}\}$.
  - Note that Riesz lets us write $\phi_w = \langle w, \cdot \rangle_{\mathcal{V}}$.
  - This is the continuous or topological dual, a subset of the algebraic dual (same definition absent 'continuous'), though these duals coincide if $\mathcal{V}$ is finite dimensional.)
  - $\mathcal{V}$ and $\mathcal{V}'$ are isometrically isomorphic.
  - Distance preserving transformation (isometry): $\|\phi_w(w)\|_{\mathcal{V}'} = \|w\|_{\mathcal{V}}$.
  - Linear bijection (isomorphism): $w \in \mathcal{V} \leftrightarrow \phi \in \mathcal{V}'$ uniquely (see [Tao09]).

Outline

Reproducing kernel Hilbert space

- Dirac delta $\delta_x : \mathcal{H} \to \mathbb{R}$ for a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$.
  - δ_x is the map from $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$.
  - For this reason it is often here called the evaluation functional.
  - δ_x is linear: $\delta_x(\alpha f + \beta g) = \alpha f(x) + \beta g(x)$.

- δ_x bounded (equiv. continuous) $\Rightarrow \delta_x = \langle \cdot, k_x \rangle_{\mathcal{H}}$ (via Riesz).

- *(Reproducing kernel Hilbert space) A Hilbert space with bounded linear evaluation functional δ_x.*

- Pause to appreciate this property: bounded δ_x means that $\exists\ k_x \in \mathcal{H}$ that achieves the action of δ_x via an inner product.
  - that is, $\delta_x f = \langle f, k_x \rangle_{\mathcal{H}} = f(x) \in \mathbb{R}$.
  - Notice the absence of any kernel in this definition.

Example and counterexample

- We have already seen $\ell_2(\mathbb{N})$ and $L_2(\mathbb{R})$; both are Hilbert spaces.
- $\ell_2(\mathbb{N})$, all countable square summable sequences:

$$\ell_2(\mathbb{N}) = \{f_i\}_{i\in\mathbb{N}}, \text{ such that } f_i \in \mathbb{R} \text{ and } \sum_{i\in\mathbb{N}} |f_i|^2 < \infty, \text{ with } \langle \{f_i\}, \{g_i\} \rangle = \sum_{i\in S} f_i g_i.$$

  - Consider $\delta_j = \langle \cdot, \mathbb{1}(i=j) \rangle_{\mathcal{H}}$ (the Kronecker delta):
  - δ_j is the evaluation operator:

$$\delta_j f = \langle f, \mathbb{1}(i=j) \rangle_{\mathcal{H}} = f_j.$$

  - δ_j is bounded (consider operator norm):

$$\|\delta_j\| = \sup_{f\in\mathcal{H}} \frac{|\delta_j f|}{\|f\|_{\mathcal{H}}} = \sup_{f\in\mathcal{H}} \frac{f_j}{\left(\sum_i |f_i|^2\right)^{\frac{1}{2}}} \leq 1 < \infty.$$

- Conclude $\ell_2(\mathbb{N})$ **is** an rkhs.

- $L_2(\mathbb{R})$, all square integrable functions (with Lebesgue measure).
  - The Dirac delta is the evaluation functional $f(x) = \int f(u)\delta(x-u)du$.
  - However, $\delta(x-u) \notin L_2(\mathbb{R})$, since $\int \delta(x-u)^2 du \not< \infty$.
- Conclude $L_2(\mathbb{R})$ **is not** an rkhs.

Reproducing kernel

- As before consider a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$.
- *(Reproducing kernel) A function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that:*

$$\begin{aligned} k_x &\triangleq k(\cdot, x) \in \mathcal{H} \qquad && \forall x \in \mathcal{X}. \\ f(x) &= \langle f, k_x \rangle_{\mathcal{H}} && \forall x \in \mathcal{X},\ \forall f \in \mathcal{H}. \end{aligned}$$

- This latter property means:
  - $\delta_x = \langle \cdot, k_x \rangle_{\mathcal{H}}$ is the evaluation functional.
  - $k_{x'}$ is also in $\mathcal{H}$, so $\delta_x k_{x'} = \langle k_x, k_{x'} \rangle_{\mathcal{H}} = k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}$.
  - ...called the *reproducing property*, as the kernel 'reproduces itself.'
- Four important (remarkable) properties follow:
  - $\mathcal{H}$ has a reproducing kernel $k \Leftrightarrow \mathcal{H}$ is an rkhs.
  - $\mathcal{H}$ has a reproducing kernel $k \Rightarrow k$ is unique.
  - Reproducing kernels k are positive definite.
  - (Moore-Aronszajn) Given a positive definite k, there exists a unique (pre-) rkhs $\mathcal{H}$ with k as its reproducing kernel.

Proof of property 1

- $\mathcal{H}$ has a reproducing kernel $k \Leftrightarrow \mathcal{H}$ is an rkhs.
- Assume $\mathcal{H}$ has a reproducing kernel k:

$$\begin{aligned} |\delta_x f| &= |\langle f, k_x \rangle_{\mathcal{H}}| \\ &\leq \|k_x\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \sqrt{\langle k_x, k_x \rangle_{\mathcal{H}}} \|f\|_{\mathcal{H}} \\ &= \sqrt{k(x,x)} \|f\|_{\mathcal{H}}. \end{aligned}$$

  ...thus δ_x is bounded, so $\mathcal{H}$ is an rkhs.

- Assume $\mathcal{H}$ is an rkhs with bounded δ_x:
  - Riesz $\Rightarrow \exists \delta_x : \delta_x f = \langle f, k_x \rangle_{\mathcal{H}} \, \forall f \in \mathcal{H}$.
  - Define a function $k(x, x') = k_x(x') \;\; \forall x, x' \in \mathbb{R}$.
  - Then $k(x, \cdot) = k_x \in \mathcal{H}$ (...first property of a reproducing kernel).
  - And $f(x) = \langle f, k_x \rangle_{\mathcal{H}}$ (...reproducing property).

  ...thus k is the reproducing kernel for $\mathcal{H}$.

Proof of property 2

- $\mathcal{H}$ has a reproducing kernel $k \Rightarrow k$ is unique.
- Assume existence of two reproducing kernels k and k'. For any $f \in \mathcal{H}$:

$$\begin{aligned} 0 &= f(x) - f(x) \\ &= \langle f, k_x \rangle_{\mathcal{H}} - \langle f, k'_x \rangle_{\mathcal{H}} \\ &= \langle f, k_x - k'_x \rangle_{\mathcal{H}} . \end{aligned}$$

Note this is enough (since $\forall f$), but the following spells it out...

- Let $f = k_x - k'_x$ (these are both in $\mathcal{H}$ so this is fine), and then:

$$\begin{aligned} \|k_x - k'_x\|^2_{\mathcal{H}} &= \langle k_x - k'_x, k_x - k'_x \rangle_{\mathcal{H}} \\ &= 0, \end{aligned}$$

... so k and k' are identical.

Proof of property 3

- Reproducing kernels k are positive definite.
- Recall we say a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if:

$$v^\top K v = \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) v_i v_j \geq 0 \quad \forall n \in \mathbb{N}_+, v \in \mathbb{R}^n.$$

- Thus:

$$\begin{aligned}
v^\top K v &= \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) v_i v_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} v_i v_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle v_i k_{x_i}, v_j k_{x_j} \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{n} v_i k_{x_i} \right\|_{\mathcal{H}}^2 \\
&\geq 0.
\end{aligned}$$

Observations

- P.D. holds for any Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$.
- Define a kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ (no reproducing property)...

$$v^\top K v = \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) v_i v_j = \ldots = \left\| \sum_{i=1}^{n} v_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0 \quad \forall n \in \mathbb{N}_+, v \in \mathbb{R}^n.$$

- All reproducing kernels are kernels with $\phi(x) = k_x$.
- We know $\exists$ non-unique feature mappings ϕ for a given kernel:

$$k(x, x') = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} x_1' \\ x_2' \\ x_1' x_2' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} x_1 \\ \frac{1}{\sqrt{2}} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sqrt{2}} x_1' \\ \frac{1}{\sqrt{2}} x_1' \\ x_2' \\ x_1' x_2' \end{bmatrix}.$$

- However, the spaces implied by the above ϕ choices are not rkhs.
- The Moore-Aronszajn theorem proves that, for every kernel k, there is a unique rkhs $\mathcal{H}$ whose reproducing kernel is k.
- Thus every kernel is the reproducing kernel of some rkhs.
- We will sketch a key piece of the proof of this theorem.

Proof sketch of property 4 (Moore-Aronszajn)

- Given a reproducing kernel k (more generally, any p.d. k), there exists a unique (pre-) rkhs $\mathcal{H}$ with k as its reproducing kernel. Define $k_x \triangleq k(\cdot, x)$.
- Construct the rkhs as the completion of the span of all k_x:

$$\mathcal{H} = \left\{ f | f = \sum_{i \in \mathbb{N}} \alpha_i k_{x_i} \quad \text{where } \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\},$$

  with inner product

$$\left\langle \sum_{i \in \mathbb{N}} \alpha_i k_{x_i}, \sum_{j \in \mathbb{N}} \alpha_j k_{x_j} \right\rangle_{\mathcal{H}} \triangleq \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \alpha_i \alpha_j k(x_i, x_j).$$

- Because k is a reproducing kernel, we have $\langle f, k_x \rangle_{\mathcal{H}} = f(x) \ \ \forall f \in \mathcal{H}$.
- Then, for a Cauchy sequence $\{f_n\}_{n \in \mathbb{N}}$ (with the fact that pointwise convergence is norm convergence in $\mathcal{H}$):

$$|f_n(x) - f(x)| = |\langle f_n - f, k_x \rangle_{\mathcal{H}}| \leq \|f_n - f\|_{\mathcal{H}} \|k_x\|_{\mathcal{H}}.$$

  ...which shows that every Cauchy sequence converges in $\mathcal{H}$ (thus complete).
- Several details omitted here; a thorough treatment is [SG12].

A few takeaways from Moore-Aronszajn

- Given a positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique (pre-) rkhs $\mathcal{H}$ with k as its reproducing kernel.
  - Every positive definite function is a reproducing kernel.
  - There is a unique rkhs $\mathcal{H}$ corresponding to each positive definite function.
  - Reminder: rkhs $\mathcal{H}$ is a subspace of functions $f : \mathcal{X} \to \mathbb{R}$; thus $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$.

Outline

Mercer's theorem

- Moore-Aronszajn placed no interesting conditions on $\mathcal{X}$ (non-empty).
- When $\mathcal{X}$ is a compact metric space (with some metric d) and k is a continuous function on that space, Mercer's theorem allows a simpler 'constructive' understanding of rkhs.
- Again $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite function.
- Fact: the integral transform $\kappa f = \int_{\mathcal{X}} k(x,u)f(u)du = g(u)$ is positive definite $\Leftrightarrow k$ is positive definite.
- Accordingly, the eigenvalues $\{\lambda_i\}$ are positive with orthonormal eigenfunctions $\phi_i : \mathcal{X} \to \mathbb{R}$:

$$\kappa\phi_i = \int_{\mathcal{X}} k(x,u)\phi_i(u)du = \lambda_i\phi_i(x).$$

cf. the more familiar discrete case.

- *(Mercer's theorem): Given the eigenvalues and eigenfunctions $\{\lambda_i, \phi_i\}$ of the integral operator defined by k, the kernel k can be written as:*

$$k(x,x') = \sum_{i\in\mathbb{N}} \lambda_i\phi_i(x)\phi_i(x'),$$

with $L_2(\mathcal{X})$ norm convergence.

Mercer's theorem

- *(Mercer's theorem): Given the eigenvalues and eigenfunctions $\{\lambda_i, \phi_i\}$ of the integral operator defined by k, the kernel k can be written as:*

$$k(x, x') = \sum_{i \in \mathbb{N}} \lambda_i \phi_i(x) \phi_i(x'),$$

  with $L_2(\mathcal{X})$ norm convergence.

- Importantly, the rkhs corresponding to this kernel k can be shown to be:

$$\mathcal{H} = \left\{ f | f = \sum_{i \in \mathbb{N}} \alpha_i \phi_i, \quad \forall \alpha_i \in \mathbb{R} \ , \ \|f\|_{\mathcal{H}} < \infty \right\},$$

  with inner product

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \sum_{i \in \mathbb{N}} \alpha_i \phi_i, \sum_{j \in \mathbb{N}} \beta_j \phi_j \right\rangle_{\mathcal{H}} \triangleq \sum_{i \in \mathbb{N}} \frac{\alpha_i \beta_i}{\lambda_i}.$$

  ... a weighted $\ell_2(\mathbb{N})$ inner product.

Why is the $\frac{1}{\lambda_i}$ factor appropriate?

- Note: $k(x,x') = \sum_{i\in\mathbb{N}} \lambda_i \phi_i(x)\phi_i(x') = \left\langle \sum_{i\in\mathbb{N}} \sqrt{\lambda_i}\phi_i, \sum_{j\in\mathbb{N}} \sqrt{\lambda_j}\phi_j \right\rangle_{L_2}$.
- Consider $f(x) = \sum_{i\in\mathbb{N}} \alpha_i \phi_i(x)$:

$$
\begin{aligned}
|f(x)|^2 &= \sum_{i\in\mathbb{N}} |\alpha_i \phi_i(x)|^2 \\
&\leq \left(\sum_{i\in\mathbb{N}} \left| \frac{\alpha_i}{\sqrt{\lambda_i}} \right|^2 \right) \left(\sum_{i\in\mathbb{N}} \left| \sqrt{\lambda_i}\phi_i(x) \right|^2 \right) \\
&= \left(\sum_{i\in\mathbb{N}} \left| \frac{\alpha_i}{\sqrt{\lambda_i}} \right|^2 \right) k(x,x),
\end{aligned}
$$

  which is finite if the sequence $\left\{ \frac{\alpha_i}{\sqrt{\lambda_i}} \right\}$ is square summable.
- Alternatively, for the reproducing property:

$$
\begin{aligned}
\langle f, k_x \rangle_{\mathcal{H}} &= \left\langle \sum_i \alpha_i \phi_i, \sum_j (\lambda_j \phi_j(x))\phi_j \right\rangle_{\mathcal{H}} \\
&= \sum_{i\in\mathbb{N}} \frac{\alpha_i \lambda_i \phi_i(x)}{\lambda_i} \\
&= f(x).
\end{aligned}
$$

Outline

Revisit sums of kernels

- Now we understand better what a kernel actually is.
- We can now return to some of our previous claims and be more rigorous.
- For example, kernel algebra:
  - We said $k = \alpha k^1 + \beta k^2$ is a kernel for $\alpha, \beta \in \mathbb{R}_+$.
  - We said $k = k^1 k^2$ is a kernel.
- The sum $k = \alpha k^1 + \beta k^2$:
  - Consider $\alpha \left\langle \phi^1(x), \phi^1(x') \right\rangle_{\mathcal{H}_1} + \beta \left\langle \phi^2(x), \phi^2(x') \right\rangle_{\mathcal{H}_2}$ in terms of all properties of an inner product:
    i. $\langle f, g \rangle = \overline{\langle g, f \rangle}$ (...which implies $\langle f, f \rangle \in \mathbb{R}$).
    ii. $\langle f, f \rangle \geq 0$,
    iii. $\langle f, f \rangle = 0 \quad \Rightarrow \quad f = 0$,
    iv. $\langle \gamma f + \rho g, h \rangle = \gamma \langle f, h \rangle + \rho \langle g, h \rangle$.
  - Essentially saying that k is positive definite if k^1, k^2 are pd and $\alpha, \beta \geq 0$.
  - If the input domains of k^1 and k^2 are the same, the resulting rkhs can be shown to be

$$\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2 = \{f_1 + f_2 : f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\},$$

with rkhs norm:

$$\|f\|^2_{\mathcal{H}} = \min_{f_1 + f_2 = f} \|f_1\|^2_{\mathcal{H}_1} + \|f_2\|^2_{\mathcal{H}_2}.$$

Roadmap

- Representer theorem.
- Kernel ridge regression.
- Posterior mean inference in a gp.
- Using the inner product $\langle f, g \rangle_{\mathcal{H}}$ (from Mercer) to understand the 'inconvenient fact' (re rkhs of a gp draw) from [Wah90, ch. 1].
- Kernel mean estimation.
- Kernel principal components analysis.
- More interesting kernel methods...

Outline

References

[Gre13] Arthur Gretton.
Introduction to rkhs, and some simple kernel algorithms.
Advanced Topics in Machine Learning. Lecture conducted from University College London, 2013.

[Hei06] Christopher Heil.
Banach and hilbert space review.
Technical Report, Georgia Tech, 2006.

[Kre89] Erwin Kreyszig.
Introductory functional analysis with applications, volume 81.
wiley New York, 1989.

[Muk15] Sayan Mukherjee.
Probabilistic machine learning.
Technical Report, Duke University, 2015.

[SC08] Ingo Steinwart and Andreas Christmann.
Support vector machines.
Springer Science and Business Media, 2008.

[SG12] Dino Sejdinovic and Arthur Gretton.
What is an rkhs?
2012.

[Tao09] Terence Tao.
245b, notes 5: Hilbert spaces.
math 245B real analysis lecture notes (https://terrytao.wordpress.com/2009/01/17/254a-notes-5-hilbert-spaces/), 2009.

[TL58] Angus Ellis Taylor and David C Lay.
Introduction to functional analysis, volume 2.
Wiley New York, 1958.

[Wah90] Grace Wahba.
Spline models for observational data, volume 59.
Siam, 1990.