# Lecture 4: Reproducing Kernel Hilbert Spaces

→ Last week: we constructed very expressive function classes

$$\mathcal{H} = \text{span}\{\sigma(\langle \cdot, w\rangle + b) \,;\, (w,b) \in \Theta\}$$

→ We verified (UAT) that these are very expressive

↳ From Fourier perspective.

Today: (·) study the framework for linear approximation in the high-dimensional limit ⇝ kernel methods

(·) relate these spaces with (shallow) NNs.

---

→ We will focus on linear models, ie functions $f_\theta : X \to \mathbb{R}$ of the form $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$ (features)

where $\varphi : X \to \mathcal{H}$ $\mathcal{H}$ a Hilbert space of very high (even infinite) dimension. $\theta \in \mathcal{H}$.

→ The key object to understand these models is the kernel function $K(x,x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$.

↠ Linear / Kernel-based models are interesting for two reasons

(i) allows for a "complete" analysis of learning (approx, statistical + optim errors can be analyzed).

(ii) enables us to better understand Neural Nets.

→ Running motivating example: $\varphi(x) = \left(\tau(w_1^T x), \dots \tau(w_m^T x)\right)$

shallow Neural Net with random weights provides random features. $w_i \sim N(0, I)$.

Q: How can we do anything practical out of an infinite-dimensional space?

## Representer Theorem

- Consider the optimization problem: given data $\{(x_i, y_i)\}_{i=1..n}$ $X \times Y$.

$$\min_{\theta \in H} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \langle \theta, \varphi(x_i)\rangle\right) + \frac{\lambda}{2}\|\theta\|_H^2.$$

→ Even though the model exists in an infinite-dim space, its effective dimension once it interacts with data is finite.

**Theorem** (Representer Thm, Kimeldorf, Wahba' 71) Let $\varphi: X \to H$ and $(x_1 \dots x_n) \in X^n$. Assume that the functional $\Psi: \mathbb{R}^{n+1} \to \mathbb{R}$ is strictly increasing along its last variable. Then

$$\inf_{\theta \in H} \Psi\left(\langle \theta, \varphi(x_1)\rangle, \dots, \langle \theta, \varphi(x_n)\rangle, \|\theta\|^2\right)$$

can be obtained by restricting $\theta$ to

$H_D := \text{span}\{\varphi(x_i)\}_{i=1\dots n}$

Proof: let $\theta \in H$. We use the fact that $H$ is Hilbert to decompose $\theta = \theta_D + \theta_\perp$ with $\theta_D \in H_D$, $\theta_\perp \in (H_D)^\perp$

Key observations: $\forall i \quad \langle \theta, \varphi(x_i)\rangle = \langle \theta_D, \varphi(x_i)\rangle$

$$\|\theta\|^2 = \|\theta_D\|^2 + \|\theta_\perp\|^2$$

$$\Psi\big(\langle\theta,\varphi(x_1)\rangle, \dots \langle\theta,\varphi(x_n)\rangle, \|\theta\|^2\big) =$$

$$= \Psi\big(\langle\theta_D,\varphi(x_1)\rangle, \dots \langle\theta_D,\varphi(x_n)\rangle, \|\theta_D\|^2 + \|\theta_\perp\|^2\big)$$

($\Psi$ is non-decreasing w.r.t last) $\geq \Psi\big(\langle\theta_D,\varphi(x_1)\rangle, \dots \langle\theta_D,\varphi(x_n)\rangle, \|\theta_D\|^2\big)$ □

→ Remark: no convexity assumption is needed!

→ Now we can write $\theta_D \in H_D$ as

$$\theta = \sum_{i=1}^n \alpha_i \varphi(x_i) \qquad \alpha \in \mathbb{R}^n$$

→ Let $K \in \mathbb{R}^{n\times n}$ $K_{ij} = k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j)\rangle_H$.

Then $\forall j$, $\langle \theta, \varphi(x_j)\rangle = \sum_i \alpha_i k(x_i, x_j) = (K\alpha)_j$

$$0 \leq \|\theta\|^2 = \Big\langle \sum_i \alpha_i \varphi(x_i), \sum_{i'} \alpha_{i'} \varphi(x_{i'})\Big\rangle = \alpha^T K \alpha \geq 0 \quad \forall \alpha$$

$K$, In particular $K$ psd + symmetric.

→ So our learning problem becomes

$$\inf_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^T K \alpha$$

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

Neither during training nor testing we need features $\varphi(x)$

only Kernel $k(x, x')$. "Kernel Trick".

Q: How to build kernels? What are some natural examples? (+ How do the resulting spaces relate to what we saw last week?)

## Kernels and RKHS (Reproducing Kernel Hilbert Space).

→ We just saw that a kernel, when it interacts with data, defines a symmetric psd kernel matrix. In fact, this is precisely what defines a pd kernel:

Def: a function $k: X \times X \to \mathbb{R}$ is a pd kernel iff all associated kernel matrices are symmetric psd.

examples: • linear kernel: $k(x, x') = x^T x'$

$$\sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \left( \sum_i \alpha_i x_i \right)^T \left( \sum_{i'=1}^{n} \alpha_{i'} x_{i'} \right)$$

$$= \left\| \sum_i \alpha_i x_i \right\|^2 \geq 0.$$

- feature map kernel: $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ with $\varphi: X \to H$ fixed.

$$\sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \left\| \sum_i \alpha_i \varphi(x_i) \right\|_H^2 \geq 0.$$

- quadratic kernel: $k(x, x') = (x^T x')^2 = \langle \varphi(x), \varphi(x') \rangle$

$$k(x, x') = x^T x' (x')^T x = \mathrm{Tr}(x^T x' (x')^T x) = \mathrm{Tr}(\underbrace{x x^T}_{\mathbb{R}^{n \times n}} \underbrace{x' (x')^T}_{\mathbb{R}^{n \times n}})$$

$$= \langle x x^T, x' x'^T \rangle \qquad \varphi(x) = x x^T.$$

→ Does the representation $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_H$ always exist? A: Yes!

**Theorem (Aronszajn '50)** $k$ is a pd kernel iff there exists a Hilbert space $H$ and a function $\varphi: X \to H$ such that $\forall x, x'$ $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_H$.

Proof: [Chapter 7 of Francis Bach's new book MCFP].

**Def/Thm: [RKHS]** Given a pd kernel, $k: X \times X \to \mathbb{R}$, there exists a unique Hilbert space $H$ with the following properties:

(i) $k(x, \cdot) : X \to \mathbb{R}$ belongs to $H$ $\forall x \in X$.

(ii) (reproducing property) function evaluation are in $H$:

$\forall x \in X, \forall f \in H \quad \langle f, k(x, \cdot) \rangle_H = f(x)$

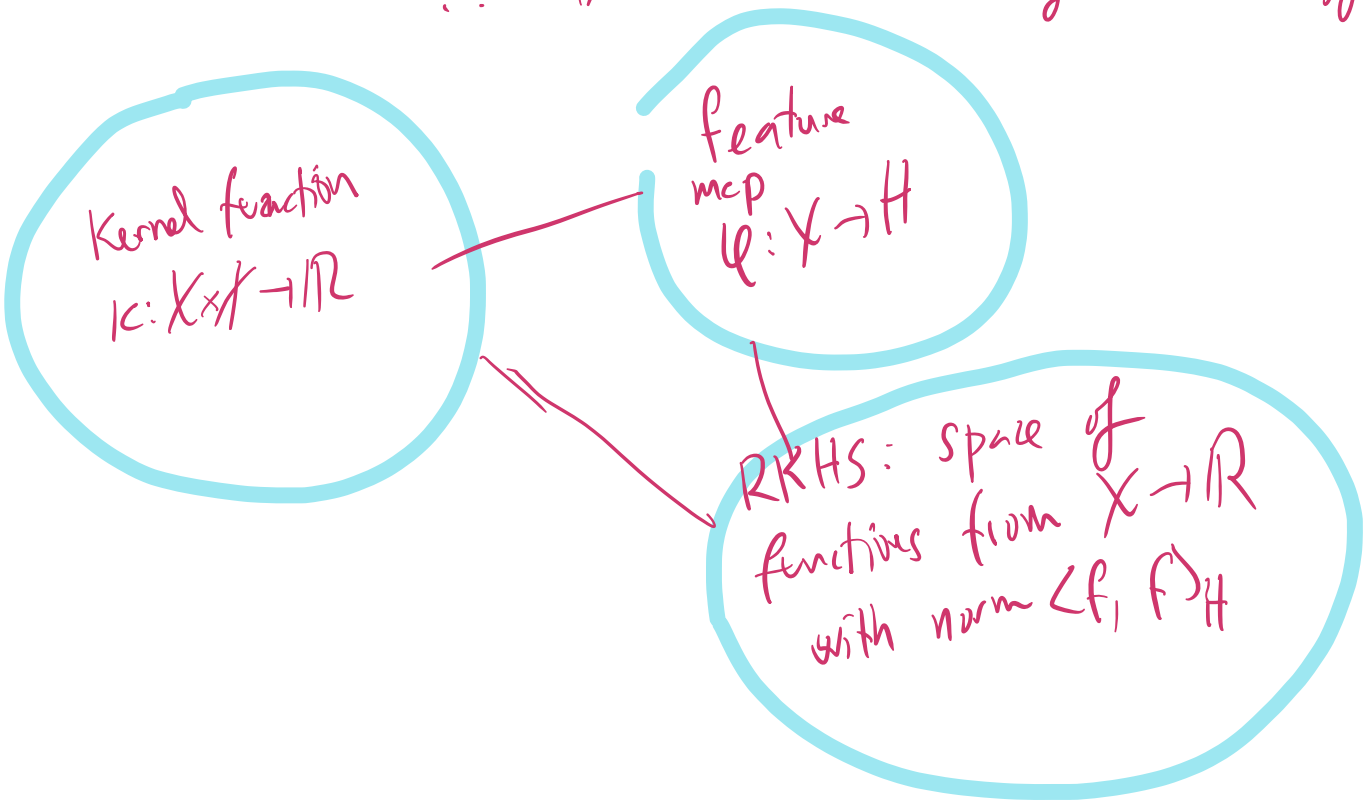
Q: Is $L_2(\mathbb{R}^d)$ an RKHS? Hell NO!

Suppose it is: then there would exist a function

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $\forall f \in L_2 \; \forall x \in \mathbb{R}^d$

$$\langle f, k(x, \cdot) \rangle = \int_{\mathbb{R}^d} f(y) k(x, y) \, dy = f(x)$$

That would imply that $k(x, y) \, dy = \delta_x$ Dirac at $x$ !! Dirac has no density wrt Lebesgue!




RKHS and Regularisation

Kernels give us a device to build hypothesis space via their associated RKHS, keeping nice Hilbertian structure.

What are the tradeoffs?

Choice of Kernel $\longleftrightarrow$ Choice of smoothness "inductive bias"

- Assume compact input space $X$.
- Let $\{\phi_i\}_{i=1\dots}$ orthonormal basis of $L_2(X)$.

Ex: $X = [0,1)$ $\{\phi_i\}$: Fourier basis ($\phi_i(t) = e^{2\pi i t}$)

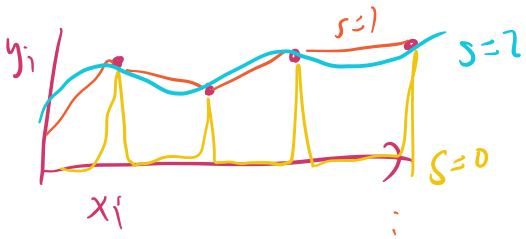
$X = S^{d-1}$ $= \{x \in \mathbb{R}^d ; \|x\|_2 = 1\}$ $\{\phi_i\}$ of spherical harmonics (basically it is the natural equivalent of Fourier w.

$f \in L^2(X)$ ; $f = \sum_{j=1}^{\infty} \alpha_j \phi_j$ $\|f\|^2_{L^2(X)} = \sum_j |\alpha_j|^2$

We already saw that regularity of $f$ can be measured in the $f$ domain by penalising this $L_2$ norm; eg

$\|f'\|^2_{L^2} = \sum_j j^2 |a_j|^2$

$\|f^{(s)}\|^2_{L^2} = \sum_j j^{2s} |a_j|^2$




$\min \|f^{(s)}\|^2_{L^2}$ st $f(x_i) = y_i$

# Penalty → Kernel ?

$f = \sum_j a_j \phi_j$ Consider a penalty

$$\sum_j b_j^2 |a_j|^2 \overset{?}{=} \|f\|_H^2$$

$$\langle f, g \rangle_H = \left( \sum_j b_j^2 \, a_j(f) \cdot a_j(g) \right)$$

$a_j(f) = \langle f, \phi_j \rangle$ , $a_j(g) = \langle g, \phi_j \rangle$

- We define $\varphi(x) = \left( b_j^{-1} \phi_j(x) \right)_{j \in \mathbb{N}}$ in $\ell_2(\mathbb{N})$

$$K(x,x') := \sum_j \frac{1}{b_j^2} \phi_j(x) \phi_j(x') \quad \left( = \langle \varphi(x), \varphi(x') \rangle_{\ell^2(\mathbb{N})} \right)$$

$K(x,x) = \sum_j 1/b_j^2 < +\infty$ $\quad \sum_j 1/b_j < +\infty$ is necessary.

$$f = \sum_j a_j \phi_j$$

$$K(x, \cdot) = \sum_j \frac{1}{b_j^2} \phi_j(x) \phi_j(\cdot)$$

$$\langle f, K(x, \cdot) \rangle_H = \sum_j b_j^2 a_j \cdot \frac{1}{b_j^2} \phi_j(x) = \sum_j a_j \phi_j(x) = f(x)$$

In the fourier case,

$K(x,x') = h(x - x')$

in particular it is translation-invariant.

in the spherical case

$K(x,x') = h(\langle x, x' \rangle)$


the kernel only depends on the angle between

$x, x'$.

* Kernel → Penalty on $L_2(X)$ ?

Define $T_K : L_2(X) \to L_2(X)$ the integral op.

$$(T_K f)(x) = \int_X K(x,y) f(y)\, dy$$

**Theorem** [Mercer] If $K$ is a pd kernel, then there exists an orthonormal basis $\{\phi_j\}_j$ st $T_K$ is diagonal, $T_K \phi_j = \lambda_j \phi_j \quad \forall j, \ \lambda_j \geq 0$.

→ As a result, $K(x,x') = \sum_j \lambda_j \phi_j(x) \phi_j(x')$

→ The RKHS can be embedded inside $L_2(X)$

$$\|f\|_H^2 = \sum_j \lambda_j^{-1} \langle f, \phi_j \rangle^2$$

Ex: $K(x,x') = h(\langle x, x' \rangle)$, $x, x' \in S^{d-1}$

$X = S^{d-1}$

$L^2(S^{d-1})$

↓

? $\{\phi_j\}_j$ $\{\lambda_j\}$ ?

↳ spherical harmonics.

$h : [-1,1] \to \mathbb{R}$

$\lambda_j \leftrightarrow$ Legendre polynomial decomposition of $h$.

→ If $K(x,x') = \mathbb{E}_{w \sim N(0,I)}\left[\sigma(w^T x)\,\sigma(w^T x')\right]$

$K(Ux, Ux') = K(x,x')$ if $U^T U = Id$.

↳ since $N(0,I)$ is Rot. inv.

$\Rightarrow K(x,x') = h(\langle x,x' \rangle)$


For $\sigma(t) = \max(0,t)$, the associated $h$ is

$$h(u) = \frac{1}{\pi}\left(u(\pi - \arccos(u)) + \sqrt{1-u^2}\right)$$

[Cho & Saul, 09].

→ For more general activations [Le Roux & Bengio]
[Francis Bach '17].

Bottom line: the smoother $h$ is, the "smaller" the associated RKHS.

## Kernels from Feature Maps

Given a feature map $\phi : X \to H$ Hilbert space

$K(x,x') = \langle \phi(x), \phi(x') \rangle_H$

$H$ is not necessarily RKHS.

Q: What is the RKHS associated with $K$?

Ex Infinite width shallow NN: $\tau$: fixed measure over $\mathbb{R}^d$.

$$K(x,x') = \mathbb{E}_{w\sim\tau}\left[\sigma(w^T x)\,\sigma(w^T x')\right]$$

$$= \int_{\mathbb{R}^d} \sigma(w^T x)\,\sigma(w^T x')\,d\tau(w)$$

$$= \langle \phi(x), \phi(x') \rangle_{L_2(\mathbb{R}^d, d\tau)}$$

$$\phi(x) = \sigma(\langle \cdot, x\rangle) \in L_2(d\tau)$$

→ There is a closely related kernel,

$$\hat{K}(x,x') = \langle \phi(x), \phi(x') \rangle_{L_2(\mathbb{R}^d, d\hat{\tau})}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \sigma(\langle x, w_i\rangle)\,\sigma(\langle x', w_i\rangle)$$

$$= \langle \underbrace{\hat{\phi}(x)}_{\mathbb{R}^m}, \hat{\phi}(x') \rangle$$

$w_1 \dots w_m \overset{iid}{\sim} \tau$

[Rahimi Recht]

$K \rightsquigarrow (H)$ RKHS over $X \to \mathbb{R}$ [Bach, 17]

$\hat{K} \rightsquigarrow \hat{H}$ a different from $H$.

Good news: as $m \to \infty$ $\hat{K} \rightsquigarrow K$

$\Rightarrow \hat{H} \to H$

**Theorem:** The RKHS of $K(x,x') = \mathbb{E}\left[\sigma(\langle x, w\rangle)\sigma(\langle w, x'\rangle)\right)$ is given by

$$\mathcal{H} = \left\{ f(x) = \int \sigma(\langle x, w\rangle)\, g(w)\, d\tau(w);\ g \in L_2(d\tau) \right\}$$

with $\|f\|_{\mathcal{H}} = \inf\left\{ \|g\|_{L^2(d\tau)};\ f(x) = \int \sigma(\langle x, w\rangle) g(w)\, d\tau(w) \right\}$

**Proof:** [Bach '17 App A].

**Remark:** $f(x) = \sigma(\langle w_\cdot, x\rangle)$ belong to the RKHS?

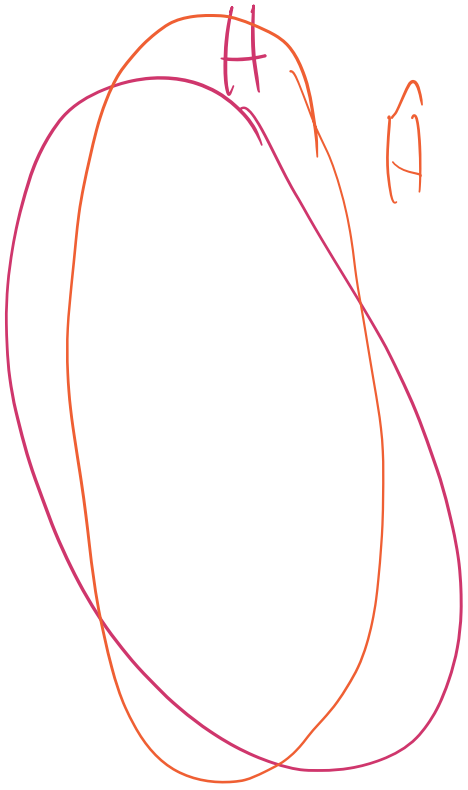
Does

No! in general ($\sigma$) we could need $g(w) = \delta_{w_\cdot}$ which is not in $L_2(d\tau)$!

$$\mathcal{H} \subseteq L^2(X)$$

$\cap$

$\mathcal{H}$




Pick $f \in \underbrace{\{f';\ \|f'\|_{\mathcal{H}} \le 1\}}_{B_1(\mathcal{H})}$

$\sup_{f \in B_1(\mathcal{H})} \|f\|_{\mathcal{H}}$