KERNEL METHODS

- Good framework for understanding linear models

$$f(x) = \langle \theta, \varphi(x) \rangle \quad \begin{cases} \theta \in \mathbb{R}^d & \underline{d \text{ very large}} \\ x \in \mathcal{X} & \underline{\text{arbitrary space}} \end{cases}$$

- Key object: <u>kernel functions</u>

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle$$

- Precise study of <u>statistical</u> & <u>approximation</u> properties in many cases

- Links to neural networks with random weights

$$\varphi(x) = \left(\sigma(w_1^T x), \dots, \sigma(w_m^T x)\right) \in \mathbb{R}^m$$

$$f(x) = \sum_{i=1}^{m} r_i \sigma(\underset{\mathcal{N}(0,1)}{w_i^T} x) = \langle r, \underline{\varphi(x)} \rangle$$

1. REPRESENTER THEOREM

$\theta \in \mathbb{R}^d$ d very large, or $d \to \infty$

Q: How can we learn such models in a tractable way

<u>ERM</u>: $$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \theta, \varphi(x_i) \rangle) + \frac{\lambda}{2} \|\theta\|_{\mathcal{H}}^2$$

$$\|\theta\|_{\mathcal{H}}^2 = \langle \theta, \theta \rangle$$

Theorem (Representer theorem, {Kimeldorf & Wahba, 1971})

Let $G(\theta) = \Psi\left(\langle\theta, \varphi(x_1)\rangle, \dots, \langle\theta, \varphi(x_n)\rangle, \|\theta\|_{\mathcal{H}}\right)$

Ψ is strictly increasing w.r.t. last variable.

Then, any minimizer of G takes the form:

$$\boxed{\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)}$$

Remark: Let $K = [\langle\varphi(x_i), \varphi(x_j)\rangle]_{ij} \in \mathbb{R}^{n\times n}$

We have $\langle\theta, \varphi(x_i)\rangle = [K\alpha]_i$

$$\|\theta\|_{\mathcal{H}}^2 = \sum_{i,j} \alpha_i \alpha_j K_{ij} = \alpha^T K \alpha$$

$\Rightarrow$ ERM becomes an optimization problem over $\alpha \in \mathbb{R}^n$

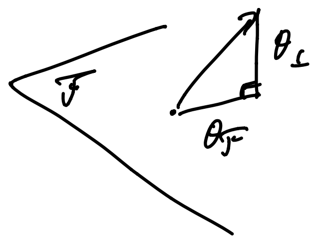
proof: Let $\mathcal{F} := \operatorname{span}\{\varphi(x_i)\}_{i=1\dots n}$

we want to show that minimizers belong to $\mathcal{F}$

Let θ be a minimizer, write

$\theta = \theta_{\mathcal{F}} + \theta_{\perp}$ with $\theta_{\mathcal{F}} \in \mathcal{F}$

$\theta_{\perp} \in \mathcal{F}^{\perp}$ (i.e. $\langle\theta_{\perp}, \varphi(x_i)\rangle = 0 \ \forall i$)



want to show $\theta_{\perp} = 0$

Assume, by contradiction, $\theta_{\perp} \neq 0$

We have $\langle \theta, \varphi(x_i) \rangle = \langle \theta_{\mathcal{F}}, \varphi(x_i) \rangle$

and $\|\theta\|^2_{\mathcal{H}} = \|\theta_{\mathcal{F}}\|^2 + \underbrace{\|\theta_{\perp}\|^2}_{> 0}$ (Pythagorean thm)

$> \|\theta_{\mathcal{F}}\|^2$

$$G(\theta) = \Psi(\langle \theta, \varphi(x_1) \rangle \dots \langle \theta, \varphi(x_n) \rangle, \|\theta\|_{\mathcal{H}})$$
$$= \Psi(\langle \theta_{\mathcal{F}}, \varphi(x_1) \rangle, \dots, \langle \theta_{\mathcal{F}}, \varphi(x_n) \rangle, \|\theta\|_{\mathcal{H}})$$
$$> G(\theta_{\mathcal{F}})$$

this contradicts θ being a minimizer.

$\Rightarrow \theta \in \mathcal{F}$

---

2. Kernels and RKHS

Def (positive-definite kernel) (p.d.)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a p.d. kernel on $\mathcal{X}$ if for any $x_1, \dots, x_n \in \mathcal{X}$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$,

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

(i.e. $K = [k(x_i, x_j)]_{ij}$ is p.s.d.)

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

$$\sum \alpha_i \alpha_j k(x_i, x_j) = \left(\sum_i \alpha_i x_i\right)^T \left(\sum_j \alpha_j x_j\right)$$
$$= \|\sum_i \alpha_i x_i\|^2 \geq 0$$

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

$$\sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \langle \sum_i \alpha_i \varphi(x_i), \sum_j \alpha_j \varphi(x_j) \rangle$$
$$= \| \sum \alpha_i \varphi(x_i) \|_H^2 \geq 0$$

- quadratic kernel $k(x,x') = (x^T x')^2 \stackrel{?}{=} \langle \phi(x), \phi(x') \rangle$

$$= x^T x' x'^T x$$
$$= Tr(x^T x' x'^T x)$$
$$= Tr(x x^T . x' x'^T)$$
$$= \langle x x^T, x' x'^T \rangle$$
$$= \langle \phi(x), \phi(x') \rangle \text{ w/ } \phi(x) = x x^T$$

$\phi(x)_{i_1 i_2 \cdots i_r} = x_{i_1} x_{i_2} \cdots x_{i_r}$

- polynomial kernel $k(x,x') = \underline{(x^T x')^r}$ (homogeneous poly.)
  $k(x,x') = (1 + x^T x')^r$ (arbitrary poly.)

Exercise:
- the sum of p.d. kernels is p.d.
- - product ______
- $k(x,x') = e^{x^T x'}$
- $k(x,x') = e^{-\alpha \|x-y\|^2}$

Theorem [Aronszajn '50]
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is p.d. if and only if there exists a Hilbert space H and $\varphi : \mathcal{X} \to H$ such that
$$k(x,x') = \langle \varphi(x), \varphi(x') \rangle_H$$

Def/Theorem (RKHS)

Let k be a p.d. kernel on $\mathcal{X}$

Then, there exist a unique Hilbert space $\mathcal{H}$ such that:

(1) $\forall x \in \mathcal{X}$, $k(x,\cdot) \in \mathcal{H}$ $\quad (k(x,\cdot) \Leftrightarrow x' \mapsto k(x,x'))$

(2) (Reproducing property)

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

$\mathcal{H}$ is called the reproducing Kernel Hibert Space (RKHS) of k.

Remark: • $\langle \underbrace{k(x,\cdot)}_{f}, k(x',\cdot)\rangle_{\mathcal{H}} = \underbrace{k(x,\cdot)}_{f}(x') = k(x,x') \Leftarrow$

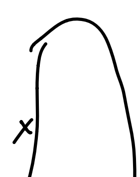
$\phi(x) = \underline{k(x,\cdot)} \in \mathcal{H} \Rightarrow k(x,x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}$

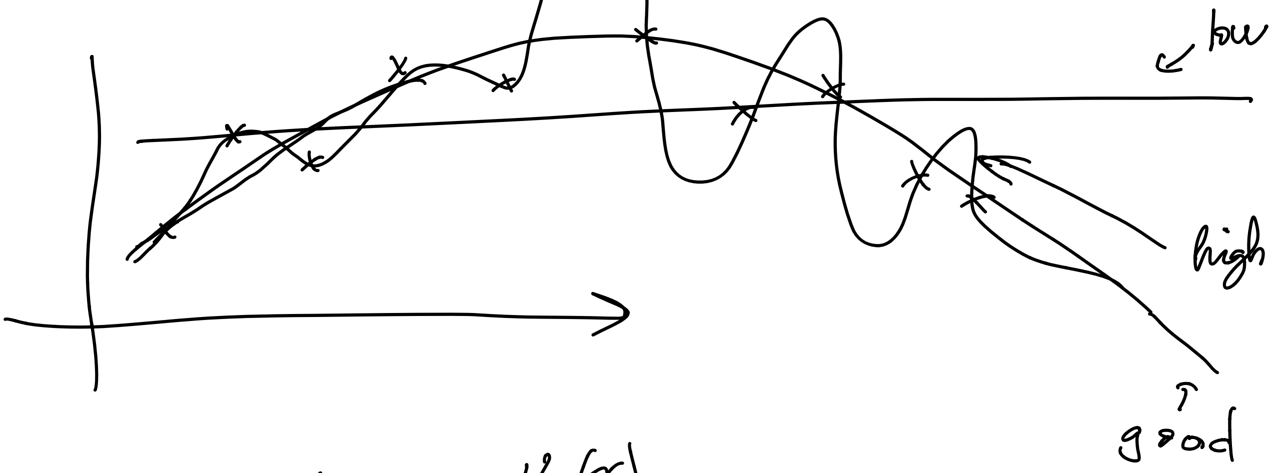
• $\langle f, \varphi(x)\rangle_{\mathcal{H}} = f(x)$ vs. $\langle \theta, \varphi(x)\rangle_H$

3. Regularization and Examples of RKHSs

Choice of kernel $\Longleftrightarrow$ choice of | norm $\|f\|^2_{\mathcal{H}}$
| regularization penalty

In ML: "prior", "inductive bias", "simplicity"





3.1 Weighted penalties in $L^2(\mathcal{X})$

Assume $\mathcal{X}$ compact

$\{\phi_i\}_i$: orthonormal basis of $L^2(\mathcal{X})$ $\rightarrow$ $\langle \phi_i, \phi_j \rangle_{L^2(\mathcal{X})} = \int_{\mathcal{X}} \phi_i(x)\,\phi_j(x)\,dx = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w.} \end{cases}$

Ex: • $\mathcal{X} = [0,1]$, ϕ_i : Fourier basis

• $\mathcal{X} = \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d,\ \|x\| = 1\}$

if $d=1$, circle $\approx [0,1]$

$d>1$ ϕ_i can be basis of Spherical Harmonics

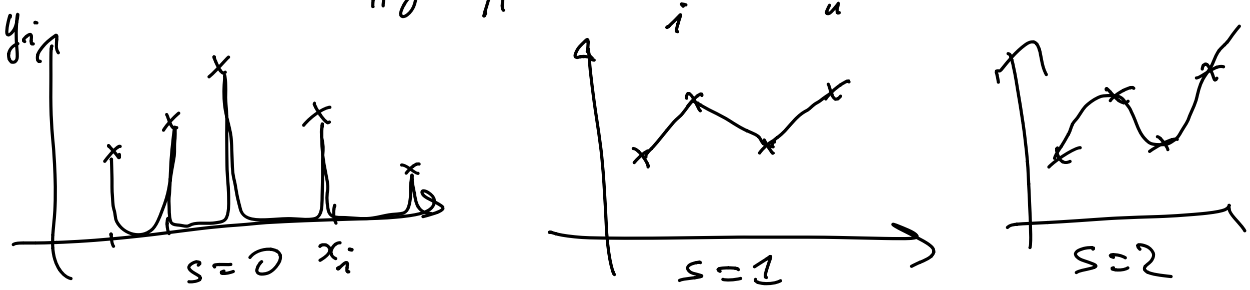
L^2 norm : $f = \sum_i a_i \phi_i$

$$\|f\|^2_{L^2(\mathcal{X})} = \sum_i a_i^2 \quad \text{(Parseval)}$$

other norms? $\|f'\|^2_{L^2(\mathcal{X})} \approx \sum_i (i a_i)^2 = \sum_i i^2 a_i^2$

$$\|f^{(s)}\|^2 \approx \sum_i i^{2s} a_i^2$$

$\min \|f^{(s)}\|^2$
s.t. $f(x_i) = y_i$



- $\underline{\text{Penalty} \to \text{Kernel ?}}$ $\qquad \langle f, g \rangle_{\mathcal{H}} = \sum_i b_i^2 a_i(f) \cdot a_i(g)$

$f = \sum_i a_i \phi_i \qquad \|f\|_{\mathcal{H}}^2 = \sum_i b_i^2 a_i^2 \qquad b_i \geq 0$

$\underline{ex}$: $\|f^{(s)}\|_{L^2}^2$: Sobolev spaces.

$b_i^2 \sim i^{2s}$

approach : . design feature map ϕ

. define corresponding kernel $k(x,x') = \langle \phi(x), \phi(x') \rangle$

. check that it yields the RKHS

Define $\phi(x) = \left(\frac{1}{b_i} \phi_i \right)_{i \in \mathbb{N}} \in \ell^2(\mathbb{N})$

$k(x,x') = \sum_i \frac{1}{\underline{b_i^2}} \phi_i(x) \phi_i(x')$ $\leftarrow$ $\underline{\text{need} : \sum_i \frac{1}{b_i^2} < \infty}$

$f = \sum a_i \phi_i$

$k(x, \cdot) = \sum_i \boxed{\frac{1}{b_i^2} \phi_i(x)} \underline{\phi_i(\cdot)}$

$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = \sum_i \cancel{b_i^2} \cdot a_i \cdot \frac{1}{\cancel{b_i^2}} \phi_i(x) = \sum_i a_i \phi_i(x) = f(x)$

$\Rightarrow$ $\mathcal{H}$ is the RKHS w/ reproducing kernel $\underline{k}$.

⚠ k is not always known in closed form.

Fourier $\Rightarrow$ $k(x,x') = \kappa(x - x')$

Spherical H $\Rightarrow$ $k(x,x') = \kappa(\langle x, x' \rangle)$ (e.g. shallow NNs)

Kernel → penalty / RKHS

Define $T_k : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ (integral operator of k)

$$T_k f(x) = \int_{\mathcal{X}} k(x,y) f(y)\, dy$$

Theorem (Mercer's thm)

If k is p.d. then there exists an orthonormal basis $\{\phi_i\}_i$ s.t. T_k is diagonal in this basis, i.e.

$$T_k \phi_i = \mu_i \phi_i \quad , \mu_i \geq 0$$

then, $k(x,x') = \sum_i \mu_i \phi_i(x) \phi_i(x')$

The RKHS of k take a similar form, with $b_i^2 = \frac{1}{\mu_i}$

$$\|f\|_{\mathcal{H}}^2 = \sum_i \frac{a_i^2}{\mu_i}$$

Ex: • $k(x,x') = \kappa(x-x')$ κ is 1-periodic, $\mathcal{X} = [0,1]$

T_k is diagonized in the Fourier basis

$\|f\|_{\mathcal{H}}^2$ penalizes high frequencies more if μ_i decays quickly

$\mu_i \leftrightarrow \hat{\kappa}_i$ (Fourier coeff of κ)

- $k(x,x') = \kappa(\langle x, x' \rangle)$, $x \in \mathbb{S}^{d-1}$

$\mu_i \Longleftrightarrow$ Legendre / Gegenbauer coefs of $\kappa : [-1,1] \to \mathbb{R}$

e.g. if $k(x,x') = \mathbb{E}_{w \sim \mathcal{N}(0,I)} [\sigma(w^T x)\sigma(w^T x')]$

then $k(Ux, Ux') = k(x,x')$ if $U^T U = I$ $\Rightarrow$ $k(x,x') = \kappa(\langle x, x' \rangle)$

(rotation - invariant)

$\to$ for $\sigma(u) = \max(u,0)$:

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

(arc-cosine kernel)

[Cho & Saul 2009]

[Bach 2017]

Convolutional Kernels: $k(x,x') = \sum_p \underbrace{k(\overset{\text{patches}}{x_p}, x_p')}_{\kappa(\langle x_p, x_p' \rangle)}$ (Mairal 2016) (Bietti-Mairal 2019)

$= \langle \phi(x), \phi(x') \rangle_H$ ← RKHS

$\phi(x) = (\bar{\varphi}(x_1), \dots, \bar{\varphi}(x_P))$

3.2. Kernels from feature maps

Feature map $\phi : \mathcal{X} \to H$, H Hilbert space (not RKHS)

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

Q: What is the RKHS for such kernels?

Ex: • sum kernel $k = k_1 + k_2$

$\varphi_1 : \mathcal{X} \to \mathcal{H}_1$ RKHS of k_1

$\varphi_2 : \mathcal{X} \to \mathcal{H}_2$ RKHS of k_2

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

$$\phi(x) = (\varphi_1(x), \varphi_2(x)) \in H = \mathcal{H}_1 \oplus \mathcal{H}_2$$

• Infinite-width NN:

$$k(x,x') = \mathbb{E}_{w \sim \tau}\{\sigma(w^T x)\sigma(w^T x')\}$$

$$= \int \sigma(w^T x)\sigma(w^T x')\, d\tau(w)$$

$$= \langle \phi(x), \phi(x') \rangle_{L^2(d\tau)}$$

$$\phi(x) = \sigma(\langle \cdot, x \rangle) \in L^2(d\tau)$$

$$k(x,x') = \frac{1}{m} \sum_{i=1}^{m} \sigma(w_i^T x)\sigma(w_i^T x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^m}$$

$\in \mathbb{R}^m$

(Random Features: [Bach 2017b]
(Rudi-Rosasco 2017))

Theorem The RKHS of $k(x,x') = \langle \phi(x), \phi(x') \rangle$ is given by

$$\mathcal{H} = \{ \langle \theta, \phi(\cdot) \rangle , \theta \in H \}$$

$$\|f\|_{\mathcal{H}}^2 = \min_{\theta \in H} \|\theta\|_H^2 \quad \text{s.t. } f = \langle \theta, \phi(\cdot) \rangle$$

(e.g. [Bach, 2017, App. A] for proof)

Ex: • (sum Kernel)

$$\|f\|_{\mathcal{H}}^2 = \min_{f_1, f_2} \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2 \quad \text{s.t. } f = f_1 + f_2$$

• (NNet) $f(x) = \int p(w) \sigma(w^T x) \, d\tau(w)$ with $p \in L^2(d\tau)$

$$\|f\|_{\mathcal{H}}^2 = \min_p \|p\|_{L^2(d\tau)}^2 \quad \text{s.t. } f(x) = \int p(w) \sigma(w^T x) \, d\tau(w)$$

finite width: $f(x) = \sum_i p_i \sigma(w_i^T x)$

⚠ $f(x) = \sigma(w_*^T x) \Rightarrow$ not always in RKHS!

(need $p(w) \rightsquigarrow$ Dirac on w_*)