Kernel method approximation

Background (Kernel ridge regression)

$$\hat{f}_\lambda = \operatorname{argmin} \left\{ \sum_{i}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_K \right\} \qquad (*)$$

$K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ PSD kernel $[K(x_i,x_j)_{ij} \succeq 0]$

$(x,y) \longmapsto K(x,y)$

$$\mathbb{K} : L^2(\mathbb{P}) \longrightarrow L^2(\mathbb{P})$$

$$\mathbb{K} g(x) = \int K(x,x') g(x') \mathbb{P}(dx')$$

$$\mathbb{K}(x_1, x_2) = \sum \lambda_e^2 \varphi_e(x_1) \varphi_e(x_2) \longleftarrow \text{ONS in } L^2(\mathbb{P})$$

$$\operatorname{Tr}(\mathbb{K}) = \int K(x,x) \mathbb{P}(dx) = \sum_{e=1} \lambda_e^2 \qquad \lambda_1^2 \geq \lambda_2^2 \geq \dots \geq \lambda_e^2$$

$$\lambda_e \to 0$$

$$\|f\|_K^2 = \sum \frac{1}{\lambda_e^2} \langle \varphi_e, f \rangle_{L^2}^2 = \langle f, \mathbb{K} f \rangle$$

<u>Example</u>: $\mathbb{P} = \operatorname{Unif}[0,1]$ $\quad f : [0, 2\pi) \to \mathbb{C} \quad \varphi_k(x) = e^{ikx}$

$$\lambda_k^2 = (1 + |k|^{2s})^{-1} \qquad s > \frac{1}{2} \qquad k \in \mathbb{Z}$$

$$\|f\|_K^2 = \sum_{q \in \mathbb{Z}} (1 + |q|^{2s}) \langle \varphi_q, f \rangle^2 = \int_0^{2\pi} \left(|f(x)|^2 + |f^{(s)}(x)|^2 \right) \frac{dx}{2\pi} \quad \text{s integer}$$

$\Longrightarrow$ encodes the smoothness of the function f.

How to solve (*): look for $f = \sum_{\ell=1}^{\infty} a_\ell \varphi_\ell$

$$\Phi = \begin{pmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots \\ \vdots & & \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots \end{pmatrix}$$

$$D = \operatorname{diag}(\lambda_1^2, \lambda_2^2, \dots)$$

$$\longrightarrow \quad \hat{a} = \operatorname{argmin} \left\{ \|y - \Phi a\|^2 + \lambda \langle a, D^{-1} a \rangle \right.$$

$$\Longrightarrow \quad \hat{a} = D \Phi^T (\lambda \mathbb{1}_n + \underbrace{\Phi D \Phi^T}_{K_n})^{-1} y$$

$$(K_n)_{ij} = \sum \lambda_\ell^2 \varphi_\ell(x_i) \varphi_\ell(x_j) = K(x_i, x_j)$$

$$\hat{f}_\lambda(x) = \sum \hat{a}_{\lambda,\ell} \varphi_\ell(x) = K(x,\cdot)^T (\lambda \mathbb{1}_n + \underbrace{K_n}_{n \times n})^{-1} y$$

$$\begin{pmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{pmatrix}$$

↳ only depends on kernel evaluated at x.

<u>Remark</u>: There is a 1-1 correspondence between choosing a kernel and choosing a featurization map

$$x \longmapsto \varphi(x) \in \mathcal{F}.$$

Back to NN: What is the kernel for (linearized) NN?

$$f(\vartheta) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_j \, \sigma(w_j \cdot x) \qquad \vartheta = \{a_j, w_j, b_j\}$$

$$m \to \infty \quad \hookrightarrow \bar{f}(\bar{\vartheta}) = f(\vartheta_0) + D_\vartheta f(\vartheta_0)(\bar{\vartheta} - \vartheta_0)$$

$\Longrightarrow$ compute

$$D_\vartheta f(\vartheta_0) = \left(\left(\partial_{a_j} f(\vartheta)\right)_j, \left(\partial_{w_j} f(\vartheta)\right)_j \right)$$

$$= \Big(\underbrace{\sigma(w_1 \cdot x), \sigma(w_2 \cdot x), \dots \sigma(w_m \cdot x)}_{\phi_{RF}(x)}, \underbrace{x a_1 \sigma'(w_1 \cdot x), \dots x a_m \sigma'(w_m \cdot x)}_{\phi_{NT}} \Big)$$

$$\begin{cases} \phi_{RF}(x) = \dfrac{1}{\sqrt{m}} \left(\sigma(w_1 \cdot x), \dots, \sigma(w_m \cdot x) \right) \\[2ex] \phi_{NT}(x) = \dfrac{1}{\sqrt{md}} \left(\sigma'(w_1 \cdot x)\, x, \dots, \sigma'(w_m \cdot x)\, x \right) \end{cases}$$

$$\Longrightarrow \bar{f}(\vartheta_0 + \gamma) = f(\vartheta_0) + \gamma_1 \cdot \phi_{RF}(x) + \sqrt{d}\, \gamma_2 \cdot \phi_{NT}(x)$$

$$\Longrightarrow \quad \Phi = \begin{pmatrix} - \phi_{RF}(x_1) - \phi_{NT}(x_1) - \\ - \phi_{RF}(x_2) - \phi_{NT}(x_2) - \\ \vdots \\ - \phi_{RF}(x_n) - \phi_{NT}(x_n) - \end{pmatrix}$$

Exercise: compute ϕ, K_m for deep network

$$\Longrightarrow \quad K_m(x_1,x_2) = \langle \phi_{RF}(x_1), \phi_{RF}(x_2) \rangle + \langle \phi_{NT}(x_1), \phi_{NT}(x_2) \rangle$$

why no a's

$$\begin{cases} K_m^{RF}(x_1,x_2) = \dfrac{1}{m} \displaystyle\sum_{j=1}^{m} \sigma(w_j \cdot x_1)\, \sigma(w_j \cdot x_2) \\ K_m^{NT}(x_1,x_2) = \dfrac{1}{md} \displaystyle\sum_{j=1}^{m} x_1 \cdot x_2 \; \sigma'(w_j \cdot x_1)\, \sigma'(w_j \cdot x_2) \end{cases}$$

These kernels converge as $m \to \infty$ to

$$\begin{cases} K^{RF}(x_1,x_2) = \mathbb{E}_w\big(\sigma(w \cdot x_1)\, \sigma(w\, x_2)\big) \\ K^{NT}(x_1,x_2) = \dfrac{x_1 \cdot x_2}{d}\, \mathbb{E}_w\big(\sigma'(w \cdot x_1)\, \sigma'(w \cdot x_2)\big) \end{cases}$$

<u>Setting</u>: $w \overset{iid}{\sim} \mathrm{Unif}(S^{d-1})$, $\quad \|x\| = \sqrt{d}$

<u>Note</u>: K^{NT}, K^{RF} are invariant under rotations

$\Longrightarrow$ can be written as functions of $\|x_1\|, \|x_2\|, x_1 \cdot x_2$

$$\Longrightarrow \exists\, h_{RF}, h_{NT} : K^{RF/NT}(x_1,x_2) = h_{RF/NT}\left(\frac{x_1 \cdot x_2}{d}\right).$$

<u>Setting</u>: $x_1, \dots x_n \overset{iid}{\sim} \mathrm{Unif}(S^{d-1}(\sqrt{d}))$

$$y_j = f_*(x_j) + \varepsilon_j \qquad \varepsilon_j \overset{iid}{\sim} N(0, \tau^2)$$

$$\mathbb{E}(f_*(x)^2) < \infty$$

Thm (Montanari et al) There exists $\lambda_0 > 0$ s.t $\forall \lambda \in [0, \lambda_0]$

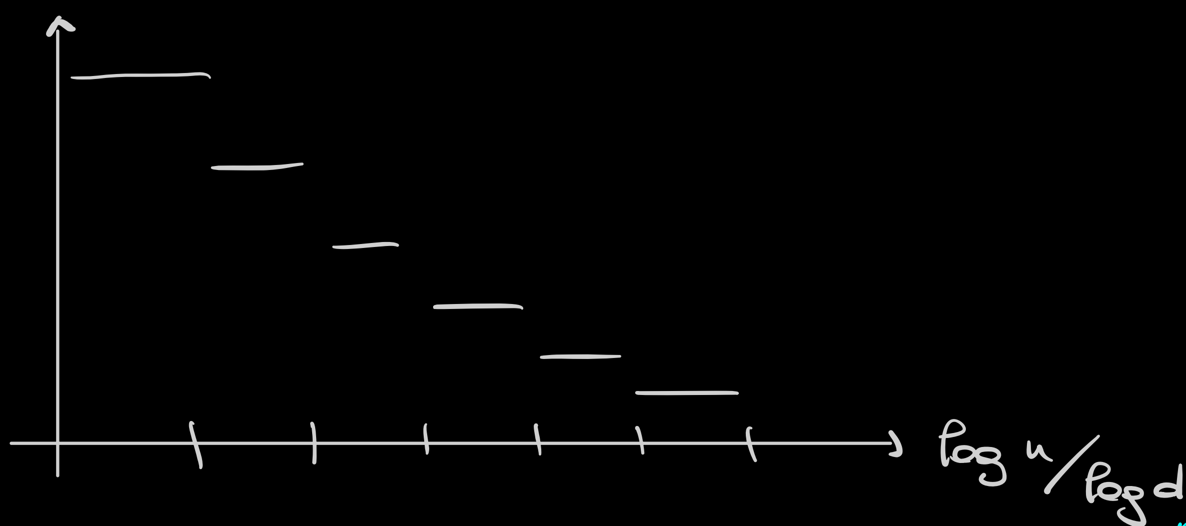
if $d^{\ell+\delta} \le n \le d^{\ell+1-\delta}$, $\delta > 0$ then

$$R(f_*, \lambda) = \mathbb{E}_x\left[\left(\hat{f}_\lambda(x) - f_*(x)\right)^2\right]$$

$$= \|P_{>\ell} f_*\|^2_{L^2(P)} + o(1)\left(\|f_*\|^2_{L^2} + \tau^2\right)$$

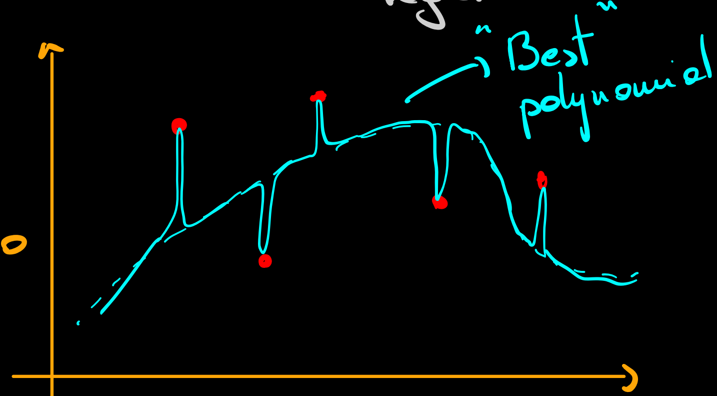
$\perp$ - orthogonal projection of f_* onto the polynomials of degree $\le \ell$. $= \perp$ - $\underset{p \in P_\ell}{\operatorname{argmin}} \|f_* - p\|^2_{L^2}$

Furthermore, no RKHS method with inner product kernel can do better



$\Longrightarrow$ · depth does not help

· benign overfitting as $\lambda \to 0$



<u>Sketch of proof</u> :

1) Diagonalizing $K(x, x')$.

ONB of $L^2(S^{d-1})$ is given by spherical harmonics

$$\mathcal{Y}_\ell = (Y_{\ell 1}, \dots Y_{\ell D(\ell)}) \qquad D(\ell) \asymp d^\ell$$

$$\int_{S^{d-1}} Y_{\ell, j} Y_{\ell', j'} \, dx = \delta_{\ell\ell'} \delta_{\ell\ell'}$$

$$P_\ell(x, x') = \sum$$

By rotational invariance we can write

$$K(x, x') = \sum_{\ell=1}^{\infty} \lambda_\ell^2 \sum_{j=1}^{D(\ell)} Y_{\ell j}$$

$$\mathrm{Tr}(K) = \int_{S^1(\sqrt{d})} K(x, x) \, dx = \sum_{\ell=0}^{\infty} \lambda_\ell^2 D(\ell) < \infty$$

$$\Longrightarrow \lambda_\ell^2 \leq D(\ell)^{-1} \lesssim d^{-\ell}$$

Population Kernel

Actually $\lambda_\ell = \frac{c_\ell}{\lambda_\ell^2}(1 + o(1))$

2) Empirical Kernel

$$K_n = \sum_{\alpha=0}^{\ell} \lambda_\alpha^2 \hat{Y}_\alpha \hat{Y}_\alpha^T + \sum_{\alpha=\ell+1}^{\infty} \lambda_\alpha^2 \hat{Y}_\alpha \hat{Y}_\alpha^T$$

$$\hat{Y}_\alpha = \underbrace{\begin{pmatrix} - & Y_\alpha(x_1) & - \\ & \vdots & \\ - & Y_\alpha(x_n) & - \end{pmatrix}}_{\longleftarrow\, D(\alpha)\, \longrightarrow} \updownarrow n$$

$\alpha \le \ell$: $\hat{Y}_\alpha$ skinny

$\rightarrow \hat{Y}_\alpha \hat{Y}_\alpha^T$ low rank

$\alpha \ge \ell+1$: $\hat{Y}_\alpha$ fat

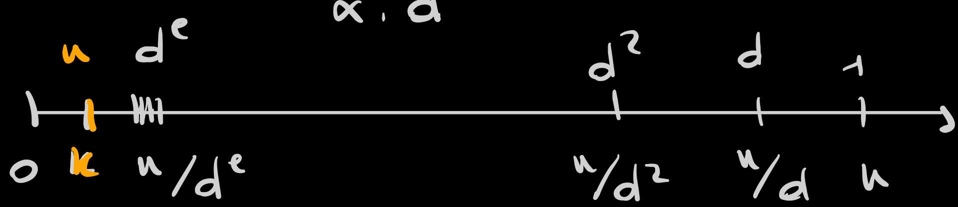
$\rightarrow \hat{Y}_\alpha \hat{Y}_\alpha^T$ full rank

$$\Longrightarrow \alpha \le \ell \quad \frac{1}{n} \hat{Y}_\alpha^T \hat{Y}_\alpha = \frac{1}{n} \sum \underbrace{Y(x_i)\, Y(i)^T}_{D(\alpha) \times D(\alpha)}$$

$$\approx \mathbb{E}\left(Y_m(x)\, Y_m(x)^T \right) \approx \mathbb{1}_{D(\alpha)}$$

$$\Longrightarrow \hat{Y}_\alpha \approx \sqrt{n}\, \hat{U}_n \qquad \hat{U}_n \text{ orthogonal}$$

$$\Longrightarrow K_n^{\le \ell} = \sum^{\ell} \frac{c_\alpha n}{\alpha!\, d^\alpha} \hat{U}_\alpha \hat{U}^T$$



On the other hand, one can prove

$$K_n^{se} = \gamma \mathbb{1}_n$$

3) Combining the above

$$\hat{f}_\lambda(x) \approx K(x,\cdot)\left(K_n^{\leq e} + (\lambda+\gamma)\mathbb{1}\right)^{-1} y$$

$$= \underbrace{K_{\leq e}\left(K_n^{\leq e} + (\lambda+\gamma)\mathbb{1}\right)^{-1} y}_{f_{\lambda,\leq e} = P_{\leq e}\hat{f}_\lambda} + \underbrace{K_{se}\left(K_n^{\leq e} + (\lambda+\gamma)\mathbb{1}\right)^{-1} y}_{f_{\lambda,se} = P_{se}\hat{f}_\lambda}$$

$$\Longrightarrow \mathbb{E}\left(\left(f_*(x) - \hat{f}_\lambda(x)\right)^2\right) =$$

$$= \underbrace{\mathbb{E}\left(\left(P_{\leq e} f_* - f_{\lambda,\leq e}\right)^2\right)}_{\text{(I) KRR in low dimensions}} + \underbrace{\mathbb{E}\left(P_{se} f_* - f_{\lambda se}\right)^2}_{\text{(II)}}$$

$$\text{(I)} \quad \mathbb{E}\left(\left(K_{\leq e}(x,\cdot)\left(K_n^{\leq e} + (\lambda+\gamma)\mathbb{1}\right)^{-1} y - P_{\leq e} f_*\right)^2\right)$$

$$y = P_{\leq e} f_* + \underbrace{\left(P_{se} f_* + \varepsilon_i\right)}_{\text{noise}} \Longrightarrow \text{KRR in } d^e \ll n$$

$$\Longrightarrow \text{(I)} = \| P_{\leq e} f_* \|^2 \cdot o_d(1) + \mathrm{Var}(\varepsilon_i) \cdot O(1)$$

$$= \left(\|f_*\|^2 + \tau^2 \right) o_d(1)$$

(II) Similarly to K_{se}

$$\| n \mathbb{E}_x \left(K_{se}(x,\cdot)^T K_{se}(x,\cdot) \right) - \gamma' I_n \| = o_d(1)$$

with $\gamma' = \sum n \lambda_k^q d^k = O_d(n d^{-(\ell+1)}) = o_d(1)$

$$\Longrightarrow \|\hat{f}_{\lambda,se}\|^2 \leq \|\mathbb{E}_x K_{se}(x,\cdot)^T K_{se}(x,\cdot)\| \, \|(K_n^{se} + (\lambda+\gamma')\mathbb{1})^{-1} y\|_2^2$$

$$\leq o_d(1) \frac{\lambda}{(\gamma' + \lambda)^2} \frac{\|y\|_2^2}{n}$$

$$\Longrightarrow (\text{II}) = \|P_{se} f_*\|_{L^2}^2 + o_d(1)$$