3 - KRR and Non-monotonic Risk Curves

- We start with generalization of kernel ridge regression.
- Give linear regression as an example and show 'double descent'.

* Kernel Ridge Regression (KRR):

- Observe n i.i.d. samples $(x_i, y_i) \sim p(x,y)$

(KRR) $$\hat{f} = \underset{f \in \underbrace{\mathcal{F}}_{\text{RKHS}}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda \|f\|_{\mathcal{F}}^2 := \hat{R}_{KRR}(f) \right\}$$

- $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$, $f: \mathcal{X} \to \mathbb{R}$ is the feature map.

$\mathcal{F}$ is an RKHS and $k(\cdot,\cdot)$ is the associated kernel.

- RKHS recap:

Def (Kernel): A kernel is a fnc $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. for any $x_1, \dots x_n \in \mathcal{X}$, the matrix $K_{ij} = k(x_i, x_j)$ is PSD.

Def (Hilbert space): A HS is an inner product space that is also a complete metric space wrt its norm.
$\mathcal{F}$ is HS, $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is inner product, which defines a norm $\|\cdot\|_{\mathcal{F}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{F}}}$

Thm ($\phi \to k$): A feature map $\phi: \mathcal{X} \to \mathcal{H}$ defines a kernel.

proof:
- $k(x, x') = \langle \phi(x), \phi(x') \rangle$
- For any $x_1 \dots x_n$ $K_{ij} = k(x_i, x_j)$ is PSD. ⊠

↳ kernel matrix $K \in \mathbb{R}^{n \times n}$

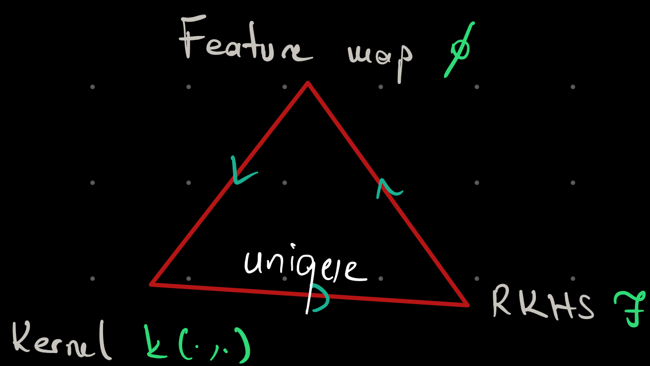
Def (RKHS): An RKHS $\mathcal{F}$ is a "nice" Hilbert space. Key property:

* Function evaluations can be written as inner products in $\mathcal{F}$

$$\forall f \in \mathcal{F}, \quad f(x) = \langle f, \psi_x \rangle_{\mathcal{F}} \quad \text{for some } \psi_x \in \mathcal{F}.$$

$\hookrightarrow$ called the representer.

Thm ($\mathcal{F} \leftrightarrow k$): Every RKHS $\mathcal{F}$ is associated with a unique kernel k.



Theorem (Representer thm): Any minimizer $\hat{f}$ of KRR is given by

$$\hat{f} = \underbrace{\sum_{i=1}^{n} \alpha_i \underbrace{k(x_i, \cdot)}_{\psi_{x_i}\text{ the representer} \longrightarrow \text{ reproducing property of } k}}_{\text{linear combination of representers.}} \quad \left\{ \text{where } \alpha = (K + n\lambda I)^{-1} Y \right\}.$$

$\rightarrow$ **Model**: $Y_i = f_*(x_i) + \varepsilon_i$ where $\varepsilon_i \perp\!\!\!\perp x_i$, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma^2$.

Sets $p(y|x)$ but nothing on $p(x)$ yet.

$$\Rightarrow \hat{R}_{KRR}(f) = \frac{1}{n}\sum_i Y_i^2 + \frac{1}{n}\sum_i f(x_i)^2 - \frac{2}{n}\sum_i Y_i f(x_i) + \lambda \|f\|_{\mathcal{F}}^2$$

Recall: Representer $\psi_x = k(x, \cdot)$ $\quad \langle f, \psi_x \rangle = f(x)$

$$= \frac{1}{n}\sum_i Y_i^2 + \frac{1}{n}\sum_i \langle f, \psi_{x_i} \rangle_{\mathcal{F}}^2 - \frac{2}{n}\sum_i Y_i \langle f, \psi_{x_i} \rangle_{\mathcal{F}} + \lambda \|f\|_{\mathcal{F}}^2$$

Define: $\hat{\Sigma} = \frac{1}{n}\sum_i \psi_{x_i} \otimes \psi_{x_i} \longrightarrow$ self-adjoint operator

$$= \frac{1}{n}\sum_i y_i^2 + \langle f, \hat{\Sigma} f\rangle_{\mathcal{F}} - 2\left\langle \underbrace{\frac{1}{n}\sum_i y_i \psi_{x_i}}_{:=\hat{b}}, f\right\rangle_{\mathcal{F}} + \lambda\langle f, f\rangle_{\mathcal{F}}$$

and define

$$= \frac{1}{n}\sum_i y_i^2 + \langle f, \hat{\Sigma} f\rangle_{\mathcal{F}} - 2\langle b, f\rangle_{\mathcal{F}} + \lambda\langle f, f\rangle_{\mathcal{F}} \quad \text{(quadratic in } f\text{)}$$

$\longrightarrow$ minimized at $\hat{f} = (\hat{\Sigma} + \lambda I)^{-1}\hat{b}$.

- We are interested in expected excess risk: $\mathbb{E}\left[\|\hat{f} - f_*\|^2_{\underset{p(x)}{L^2(p)}}\right]$

$$\mathbb{E}\left[\|\hat{f} - f_*\|^2_{L^2(p)}\right] = \mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i y_i \psi_{x_i} - f_*\right\|^2_{L^2(p)}\right]$$

$\hookrightarrow = f_*(x_i) + \varepsilon_i$

$= \langle \psi_{x_i}, f_*\rangle + \varepsilon_i$ (!!) can we do this?

$$= \mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \psi_{x_i}\{\langle\psi_{x_i}, f_*\rangle + \varepsilon_i\} - f_*\right\|^2_{L^2(p)}\right]$$

$\varepsilon_i \perp\!\!\!\perp x_i$

$$= \mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \varepsilon_i \psi_{x_i}\right\|^2_{L^2(p)}\right] + \mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \psi_{x_i}\langle\psi_{x_i}, f_*\rangle - f_*\right\|_{L^2(p)}\right]$$

Variance $\triangleq V(\lambda)$ Bias $\triangleq B(\lambda)$

Let $\Sigma = \mathbb{E}\hat{\Sigma}$ (or $\mathbb{E}[\psi_x \otimes \psi_x]$) and observe

(!) $\|g\|^2_{L^2(p)} = \int g(x)^2 \, dp(x) = \int \langle g, \psi_x\rangle^2_{\mathcal{F}} \, dp(x) = \left\langle g, \int \psi_x \otimes \psi_x \, dp(x) \, g\right\rangle_{\mathcal{F}}$

$= \langle g, \Sigma g\rangle_{\mathcal{F}}$.

$$V(\lambda) \overset{\text{by (!)}}{=} \mathbb{E}\left[\left\langle (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \varepsilon_i \psi_{x_i}, \Sigma(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \varepsilon_i \psi_{x_i}\right\rangle_{\mathcal{F}}\right]$$

$$= \frac{1}{n^2}\mathbb{E}\left[\mathrm{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\underbrace{\sum_i \varepsilon_i^2 \varphi_{x_i}\otimes\varphi_{x_i}}_{\mathbb{E}_\varepsilon \to \,=\hat{\Sigma}\cdot n\cdot\sigma^2}\right)\right]$$

(V) $$= \frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma\underbrace{(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}}_{\preceq I}\right)\right]$$

since $(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} \preccurlyeq I$

$$\le \boxed{\frac{\sigma^2}{n}\mathbb{E}\left[\mathrm{Tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma\right)\right]}$$

To simplify calculations, we assume $f_* \in \mathcal{F} \Rightarrow \langle f_*, \varphi_x\rangle = f_*(x)$

$B(\lambda)$: $$\mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_i \varphi_{x_i}\langle\varphi_{x_i}, f_*\rangle - f_*\right\|^2_{L^2(p)}\right]$$

$$= \mathbb{E}\left[\left\|(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}f_* - f_*\right\|^2_{L^2(p)}\right]$$

by (!)

$$= \mathbb{E}\left[\left\langle\left\{(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}-I\right\}f_*, \Sigma\left\{(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}-I\right\}f_*\right\rangle_{\mathcal{F}}\right]$$

$$(\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}-I = (\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma}+\lambda I-\lambda I) - I$$
$$= I - \lambda(\hat{\Sigma}+\lambda I)^{-1} - I$$

(B) $$\boxed{= \mathbb{E}\left[\left\|\lambda\Sigma^{\frac{1}{2}}(\hat{\Sigma}+\lambda I)^{-1}f^*\right\|^2_{\mathcal{F}}\right]}$$

- $\hat{\Sigma}$ concentrates around Σ (Bach p188) for constant d. Need $\|\varphi_x\|_{\mathcal{F}} \le R$
- First term $\approx \frac{\sigma^2}{n}\mathrm{Tr}\left((\Sigma+\lambda I)^{-1}\Sigma\right) = O\left(\frac{\sigma^2}{n\lambda}\right)$
- Second term $\approx \lambda^2\left\langle f_*, \underbrace{(\Sigma+\lambda I)^{-1}\Sigma}_{\le I}\underbrace{(\Sigma+\lambda I)^{-1}}_{\le \frac{1}{\lambda}}f_*\right\rangle_{\mathcal{F}} = O\left(\lambda\|f_*\|^2_{\mathcal{F}}\right)$

$$\Rightarrow \quad \text{Excess Risk} \lesssim \frac{\sigma^2}{n\lambda} + \lambda\|f^*\|^2_{\mathcal{F}}$$

- choosing $\lambda = \frac{1}{\sqrt{n}}$ $\lesssim \frac{1}{\sqrt{n}} \Rightarrow$ generalization!

Theorem: Let $y_i = f^*(x_i) + \varepsilon_i$ for $i = 1 \dots n$ for $f^* \in \mathcal{F}$ and $\hat{f} = \underset{\mathcal{F}}{\operatorname{argmin}} \hat{R}_{KRR}(f)$ for $\lambda = \frac{1}{\sqrt{n}}$. Then, if $\|\varphi_x\|_{\mathcal{F}} \leq R \ \forall x$, we have

$$\mathbb{E}\left[\|\hat{f} - f_*\|^2_{L^2(p)}\right] \lesssim \frac{1}{\sqrt{n}}.$$

- Double descent in linear regression

- RKHS: $\mathcal{F} = \left\{ f_\theta(x) = \langle \theta, x \rangle : \theta \in \mathbb{R}^d \right\}$

- $\langle \cdot, \cdot \rangle_{\mathcal{F}}$: $\langle f_\theta, f_w \rangle_{\mathcal{F}} = \langle \theta, w \rangle_{\mathbb{R}^d} = \theta^T w$.

- Representer: $\langle \varphi_x, f_\theta \rangle_{\mathcal{F}} = f_\theta(x) = \langle \theta, x \rangle$

$\varphi_x(y) = \langle x, y \rangle$ (or $\varphi_x = f_x$)

- Model: $Y = \langle \theta_*, x \rangle + \varepsilon$ $\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$\left.\begin{array}{l} x \sim \mathcal{N}(0, I) \\ \theta_* \sim \mathcal{N}(0, \frac{1}{d} I) \end{array}\right\} \ \mathbb{E}\langle \theta_*, x \rangle^2 = 1$$

- $$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \theta, x_i \rangle)^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

$$= (\underbrace{\tfrac{1}{n} X^T X}_{\hat{\Sigma}} + \lambda I)^{-1} \frac{1}{n} X^T y$$

- Excess Risk: $\mathbb{E}\left[\|\hat{\theta} - \theta_*\|^2\right] \triangleq ER(\lambda)$

By (V) and (B) $\boxed{ER(\lambda) = B(\lambda) + V(\lambda)}$ where

$$B(\lambda) = \lambda^2\, \mathbb{E}\left[\left\langle \theta^*, (\hat{\Sigma}+\lambda I)^{-2}\theta_* \right\rangle\right] \qquad V(\lambda) = \frac{\sigma^2}{n}\, \mathbb{E}\left[\mathrm{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-2}\hat{\Sigma}\right)\right]$$

$$= \frac{\lambda^2}{d}\, \mathbb{E}\left[\mathrm{Tr}\left(\left(\hat{\Sigma}+\lambda I\right)^{-2}\right)\right] \qquad = \sigma^2 \frac{d}{n}\, \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} \frac{\lambda_i}{(\lambda_i+\lambda)^2}\right]$$

$$= \lambda^2\, \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} \frac{1}{(\lambda_i+\lambda)^2}\right]$$

where λ_i's are the eigenvalues of $\hat{\Sigma}$.

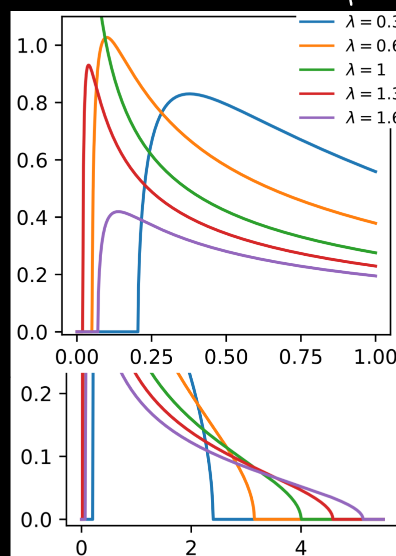
— Marchenko – Pastur Law: Let $d, n \to \infty$ and $\frac{d}{n} \to \gamma$.

Let $X \in \mathbb{R}^{n\times d}$ s.t. X_{ij} are iid mean, variance 1.

Then, for any reasonable fnc ϕ and $\hat{\Sigma} = \frac{1}{n}X^T X$

$$\frac{1}{d}\sum_{j=1}^{d} \phi\left(\lambda_j(\hat{\Sigma})\right) \xrightarrow{a.s.} \int \phi \, d\mu$$

$\to \gamma$



wikipedia

where μ is the M-P law given as

$$\frac{d\mu}{dx} = \begin{cases} (1-\gamma^{-1})\,\delta_0(x) + \nu(x) & \text{if } \gamma > 1 \\ \nu(x) & \text{if } \gamma \in [0,1] \end{cases}$$

and
$$\nu(x) = \begin{cases} \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - x)(x-\gamma_-)}}{\gamma x} & x \in [\gamma_-, \gamma_+] \\ 0 & \text{otw} \end{cases}$$

with $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$

* Stieltjes transform: of M-P law:

$$s(z) = \int \frac{1}{x-z}\, d\mu(x) \qquad s(-z) = \frac{-1+\gamma-z+\sqrt{(1-\gamma+z)^2+4\gamma z}}{2\gamma z} \quad \text{for } z>0.$$

* For linear regression: $ER(\lambda) = V(\lambda) + B(\lambda)$

V :
$$V(\lambda) = \sigma^2 \frac{d}{n}\, \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d} \frac{\lambda_i}{(\lambda_i+\lambda)^2}\right]$$

$$\longrightarrow \sigma^2 \gamma \int \frac{x}{(\lambda+x)^2}\, d\mu(x) = \sigma^2\gamma \left\{ \int \frac{1}{x+\lambda}\, d\mu(x) - \int \frac{\lambda}{(\lambda+x)^2}\, d\mu(x) \right\}$$

$$= \sigma^2 \gamma \left\{ s(-\lambda) - \lambda s'(-\lambda) \right\}.$$

B: $$B(\lambda) \longrightarrow \lambda^2 \int \frac{1}{(\lambda+x)^2} d\mu(x) = \lambda^2 s'(-\lambda)$$

Theorem: Let $Y_i = \langle \theta_*, x_i \rangle + \varepsilon_i$ for $x \sim N(0, I) \perp\!\!\!\perp \varepsilon \sim N(0, \sigma^2)$ and $\theta_* \sim N(0, \frac{1}{d} I)$. Then, the ridge regression solution

$$\hat{\theta}_\lambda = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \theta, x_i \rangle)^2 + \frac{\lambda}{2} \|\theta\|_2^2$$

satisfies $\mathbb{E}\left[\|\hat{\theta}_\lambda - \theta_*\|^2\right] \triangleq ER(\lambda) = B(\lambda) + V(\lambda)$ where as $\frac{d}{n} \to \gamma$

$$B(\lambda) \to \lambda^2 s'(-\lambda)$$
$$V(\lambda) \to \sigma^2 \gamma \left\{ s(-\lambda) - \lambda s'(-\lambda) \right\}$$ almost surely.

Remarks:

- "Ridgeless" case $(\lambda \downarrow 0) \Rightarrow$ Minimum norm solution when $d > n$. Gradient descent can find this! Implicit regularization (A1)

$$\lim_{\lambda \downarrow 0} B(\lambda) = B(0_+) = \lim_{\lambda \downarrow 0} \lambda^2 s'(-\lambda) = \begin{cases} 0 & \gamma < 1 \\ 1 - \frac{1}{\gamma} & \gamma \geq 1 \end{cases}$$

(A1)

$$\lim_{\lambda \downarrow 0} V(\lambda) = V(0_+) = \lim_{\lambda \downarrow 0} \sigma^2 \gamma \left\{ s(-\lambda) - \lambda s'(-\lambda) \right\} = \sigma^2 \begin{cases} \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \frac{1}{\gamma - 1} & \gamma \geq 1 \end{cases}$$

* Variance diverges only when $\lambda = 0_+$.