Theorem 7 $\frac{1}{n}\mathbb{E}\sum_{i=1}^n (f^0(x_i)-\hat f_\lambda(x_i))^2 \le \underbrace{\frac{\sigma^2}{n}\frac{1}{\lambda}\sum_{i=1}^n \min\left(\frac{d_i}{4},\lambda\right)}_{\text{variance}} + \underbrace{\frac{\lambda}{4n}}_{\text{bias}}$

Proof: We know from the representer theorem

$$(\hat f_\lambda(x_1),\dots \hat f_\lambda(x_n))^T = K(K+\lambda I)^{-1}Y$$

Also $(f^0(x_1),\dots,f^0(x_n))^T = K\alpha$ for some $\alpha\in\mathbb{R}^n$ and $1\ge \|f^0\|_{\mathcal{H}}^2 \ge \alpha^T K\alpha$ (see ex sheet)

Consider the eigendecomposition $K=UDU^T$ with $D_{ii}=d_i$.

We have

$$\mathbb{E}\sum_{i=1}^n (f^0(x_i)-\hat f_\lambda(x_i))^2 = \mathbb{E}\|K\alpha - K(K+\lambda I)^{-1}(K\alpha+\varepsilon)\|_2^2$$

$$= \mathbb{E}\|\cancel{U}DU^T\alpha - \cancel{U}DU^T\underbrace{(UDU^T+\lambda I)^{-1}}_{U(D+\lambda I)^{-1}U^T}(UDU^T\alpha+\varepsilon)\|_2^2$$

$$= \mathbb{E}\|DU^T\alpha - D(D+\lambda I)^{-1}DU^T\alpha - D(D+\lambda I)^{-1}U^T\varepsilon\|_2^2$$

$$= \underbrace{\|\underbrace{DU^T\alpha}_{=\theta} - D(D+\lambda I)^{-1}\underbrace{DU^T\alpha}_{\theta}\|_2^2}_{1} + \underbrace{\mathbb{E}\|D(D+\lambda I)^{-1}U^T\varepsilon\|_2^2}_{2}$$

$1 = \|\{I-D(D+\lambda I)^{-1}\}\theta\|_2^2$

ith diagonal element of $I-D(D+\lambda I)^{-1}$ is $1-\frac{d_i}{d_i+\lambda} = \frac{\lambda}{d_i+\lambda}$

So $1 = \sum_{i=1}^n \frac{\lambda^2}{(d_i+\lambda)^2}\theta_i^2$

Now $1\ge \alpha^T K\alpha = \alpha^T UDU^T\alpha$

$= \alpha^T UDD^+DU^T\alpha$ where D^+ is diagonal and $D_{ii}^+ = \begin{cases} d_i^{-1} & \text{if } d_i>0 \\ 0 & \text{o/w}\end{cases}$

$= \sum_{i:d_i>0}\frac{\theta_i^2}{d_i}$

$$\sum_{i=1}^n \frac{\lambda^2}{(d_i+\lambda)^2}\theta_i^2 = \sum_{i:d_i>0}\frac{\theta_i^2}{d_i}\frac{d_i\lambda^2}{(d_i+\lambda)^2} \quad (\text{when } d_i=0,\ \theta_i=0)$$

$$\le \lambda\max_{i=1,\dots,n}\frac{d_i\lambda}{(d_i+\lambda)^2} \le \frac{\lambda}{4} \quad \left[(a+b)^2\ge 4ab\right]$$

$$② = \mathbb{E}\,\mathrm{tr}\left(\varepsilon^T U D (D+\lambda I)^{-2} D U^T \varepsilon\right)$$
$$= \mathbb{E}\,\mathrm{tr}\; D(D+\lambda I)^{-1} U^T \varepsilon\varepsilon^T U D (D+\lambda I)^{-1}$$
$$= \sigma^2\,\mathrm{tr}\left(D^2(D+\lambda I)^{-2}\right) = \sigma^2 \sum_{i=1}^n \frac{d_i^2}{(d_i+\lambda)^2}$$

$$\frac{d_i^2}{(d_i+\lambda)^2} = \frac{d_i}{\lambda}\,\frac{d_i\lambda}{(d_i+\lambda)^2} \le \min\left(\frac{d_i}{4\lambda}, 1\right) \qquad \square$$

<u>Not examinable</u>:

If the x_i are now random and iid and independent of ε_i, our previous analysis holds by conditioning on the $x_1, \dots, x_n$.

Define $\hat{\mu}_i = \frac{d_i}{n}$, $\lambda_n = \frac{\lambda}{n}$. Have

$$\frac{1}{n}\mathbb{E}\sum_{i=1}^n \left(f^0(x_i) - \hat{f}_\lambda(x_i)\right)^2 \le \frac{\sigma^2}{\lambda_n}\,\frac{1}{n}\sum_{i=1}^n \mathbb{E}\min\left(\frac{\hat{\mu}_i}{4}, \lambda_n\right) + \frac{\lambda_n}{4} = \mathbb{E}\,\delta_n(\lambda_n)$$

To relate $\mathbb{E}\min\left(\frac{\hat{\mu}_i}{4}, \lambda_n\right)$ more directly to k, we can use Mercer's theorem.

Given a density $p(x)$ on $\mathcal{X}$, under technical conditions, we have

$$k(x, x') = \sum_{j=1}^\infty \mu_j e_j(x) e_j(x') \quad \forall x, x' \in \mathcal{X}$$

where $e_j \in \mathcal{H}$ are eigenfunctions and $\mu_j \ge 0$ are eigenvalues obeying

$$\mu_j e_j(x') = \int_{\mathcal{X}} k(x, x') e_j(x) p(x)\,dx$$

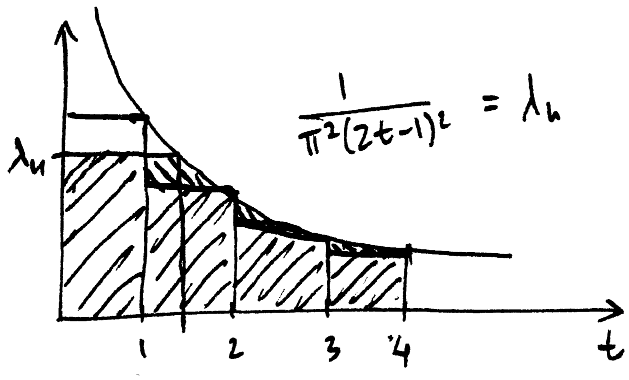
and $\int_{\mathcal{X}} e_j(x) e_k(x) p(x)\,dx = \mathbb{1}_{\{j=k\}}$.

Can show that (up to constant multiples)

$$\frac{1}{n}\mathbb{E}\sum_{i=1}^n \min\left(\frac{\hat{\mu}_i}{4}, \lambda_n\right) \le \frac{1}{n}\sum_{i=1}^\infty \min\left(\frac{\mu_i}{4}, \lambda_n\right)$$

and when k is the Sobolev kernel and $p(x)$ is the uniform density on $[0,1]$

$$\frac{\mu_j}{4} = \frac{1}{\pi^2(2j-1)^2}$$



$$\frac{1}{\pi^2(2t-1)^2} = \lambda_n \qquad t = \frac{1}{2}\left(\frac{1}{\sqrt{\pi^2\lambda_n}} + 1\right)$$

$$\sum_{i=1}^{\infty} \min\left(\frac{\mu_i}{4}, \lambda_n \right) \le \frac{\lambda_n}{2}\left(\frac{1}{\sqrt{\pi^2 \lambda_n}} + 1 \right) + \frac{1}{\pi^2} \int_{\frac{1}{2}\left(\frac{1}{\sqrt{\pi^2\lambda_n}}+1\right)}^{\infty} \frac{1}{(2t+1)^2}\, dt$$

$$= \frac{\sqrt{\lambda_n}}{2\pi} + \frac{\lambda_n}{2} + \frac{1}{\pi^2}\,\frac{1}{2}\sqrt{\pi^2 \lambda_n}$$

$$= \frac{\sqrt{\lambda_n}}{\pi} + \frac{\lambda_n}{2}$$

Therefore as $n \to \infty$,

$$\mathbb{E}\, \mathcal{J}_n(\lambda_n) = \mathcal{O}\left(\frac{\sigma^2}{n\sqrt{\lambda_n}} + \lambda_n \right)$$

so for an optimal $\lambda_n \sim \left(\frac{\sigma^2}{n}\right)^{2/3}$ we get an error rate of $\left(\frac{\sigma^2}{n}\right)^{2/3}$.

TEoSL 129-135
417-438