Examples

Linear kernel

$$\mathcal{H} = \{ f : f(x) = \sum_{j=1}^{n} \alpha_j x_j^T x \text{ for some } n \in \mathbb{N}, \alpha \in \mathbb{R}^n, x_1, \ldots, x_n \in \mathbb{R}^p \}$$

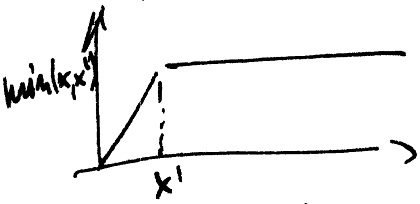
or $\mathcal{H} = \{ f : f(x) = \beta^T x \text{ for some } \beta \in \mathbb{R}^p \}$

If $f(x) = \beta^T x$, i.e. $f(\cdot) = k(\cdot, \beta)$, then $\|f\|_{\mathcal{H}}^2 = \langle k(\cdot, \beta), k(\cdot, \beta) \rangle$
$= k(\beta, \beta) = \|\beta\|_2^2$

Sobolev kernel

$k(x, x') = \min(x, x')$, $x, x' \in [0, 1]$

$\mathcal{H}$ includes linear combinations of functions $x \longmapsto \min(x, x')$ for $x' \in [0,1]$ and their pointwise limits.



Can show that $\mathcal{H}$ includes all Lipschitz functions on $[0,1]$ that are 0 at the origin (Lipschitz means $|f(x) - f(x')| \leq L|x - x'|$ for some L $\forall x, x' \in [0,1]$).

The norm $\left(\int_0^1 f'(x)^2 \, dx \right)^{1/2}$

e.g. $\|\min(\cdot, x')\|_{\mathcal{H}}^2 = \min(x', x') = x'$

Also $\int_0^{x'} 1^2 \, dx = x'$ ✓

1.4.3 The representer theorem

What optimisation problem is kernel ridge regression (KRR) solving? An alternative way of writing the usual ridge optimisation is

$$\underset{f \in \mathcal{H}}{\arg\min} \Big\{ \sum_{i=1}^{n} (Y_i - \underbrace{f(x_i)}_{x_i^T \beta})^2 + \lambda \underbrace{\|f\|_{\mathcal{H}}^2}_{\|\beta\|_2^2} \Big\}$$

where $\mathcal{H}$ is the RKHS of the linear kernel.

Theorem 6 (the representer theorem)

Let c be an arbitrary loss function and suppose J is strictly increasing. Let $\mathcal{H}$ be an RKHS with reproducing kernel k. Then $\hat{f} \in \mathcal{H}$ minimises

$$Q_1(f) = c(Y, X, f(x_1), \ldots, f(x_n)) + J(\|f\|_{\mathcal{H}}^2) \text{ over } f \text{ iff}$$

$$\hat{f}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_i k(\cdot, x_i) \text{ and } \hat{\alpha} \in \mathbb{R}^n \text{ minimises}$$

$$Q_2(\alpha) = c(Y, X, K\alpha) + J(\underbrace{\alpha^T K \alpha}_{\hat{\alpha}^T K \hat{\alpha} = \|\hat{f}\|_{\mathcal{H}}^2}) \text{ over } \alpha \in \mathbb{R}^n.$$

Remark Specialising to ridge, (*) is equivalent to minimising $\|Y - K\alpha\|_2^2 + \alpha^T K \alpha$. Example 1 qu 9 shows that minimiser $\hat{\alpha}$ satisfies $K\hat{\alpha} = K(K+\lambda I)^{-1} Y$. Thus KRR is exactly (*).

Proof Suppose $\hat{f}$ minimises Q_1. We may write $\hat{f} = u + v$ where
$u \in V = \text{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\}$ and $v \in V^\perp$.

Then $\hat{f}(x_i) = \langle \hat{f}, k(\cdot, x_i)\rangle = \langle u + \cancel{v}^{\,0}, k(\cdot, x_i)\rangle = u(x_i)$

Meanwhile $J(\|f\|_{\mathcal{H}}^2) = J(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$ with equality iff $v = 0$.

By optimality of $\hat{f}$, $v = 0$. So $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$, and again by optimality of $\hat{f}$, $\hat{\alpha}$ must minimise Q_2.

Now suppose $\hat{\alpha}$ minimises Q_2 and set $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$
if $\tilde{f} \in \mathcal{H}$ with $Q_1(\tilde{f}) \leq Q_2(\hat{f})$, by the argument above, we can write $\tilde{f} = u + v$ with $u \in V$, $v \in V^\perp$ and we know $Q_1(u) \leq Q_1(\tilde{f})$. But by optimality of $\hat{f}$, $Q_1(\hat{f}) \leq Q_1(u)$ so $Q_1(\hat{f}) = Q_1(\tilde{f})$. $\square$

The representer theorem gives the form of the entire fitted regression $\hat{f}$ (not just the fitted vals).
e.g. with KRR, given a new obs x, our prediction would be

$$\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i k(x, x_i)$$

1.5 Kernel ridge regression

Consider a model $Y_i = f^0(x_i) + \varepsilon_i$, $\text{Var}(\varepsilon) = \sigma^2 I$, $\mathbb{E}\varepsilon = 0$.
Assume $f^0 \in \mathcal{H}$ where $\mathcal{H}$ is a RKHS with reproducing kernel k. Assume $\|f^0\|_{\mathcal{H}} \leq 1$.
Let K be the kernel matrix $K_{ij} = k(x_i, x_j)$ with eigenvalues $d_1 \geq d_2 \geq \dots \geq d_n \geq 0$
Define $\hat{f}_\lambda = \underset{f \in \mathcal{H}}{\arg\min} \{\sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2\}$.

Theorem 7 The mean squared prediction error of $\hat{f}_\lambda$ (MSPE) is

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