

1 Kernel Machines

Consider a linear model

$$Y_i = x_i^T \beta^0 + \varepsilon_i, \quad i = 1, \dots, n, \quad x_i \in \mathbb{R}^p \text{ fixed}$$

where $\mathbb{E}\varepsilon = 0$, $\text{Var}(\varepsilon) = \sigma^2 I_n$. We have

$$\begin{aligned} \hat{\beta}^{\text{ols}} &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^n (Y_i - x_i^T \beta)^2 \\ &= \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|^2 \\ &= (X^T X)^{-1} X^T Y. \end{aligned}$$

Classical theory:

- $\hat{\beta}^{\text{ols}}$ unbiased,

$$\text{Var}(\hat{\beta}^{\text{ols}}) = \sigma^2 (X^T X)^{-1} = i^{-1}(\beta^0)$$

Where i is the Fisher information.

- Cramér-Rao lower bound: if an estimator $\tilde{\beta}$ is unbiased then

$$\text{Var}(\tilde{\beta}) - i^{-1}(\beta^0) \underset{\text{positive semi-definite}}{\geq} 0.$$

- If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then $\hat{\beta}^{\text{ols}}$ is the MLE of β^0 . Furthermore $\sqrt{n}(\hat{\beta}^{\text{ols}} - \beta^0) \sim \mathcal{N}(0, n\sigma^2(X^T X)^{-1})$. From this we can derive confidence intervals, hypothesis test, etc.

In a general model with parameter $\theta \in \mathbb{R}^p$, n independent observations, under regularity, we have asymptotic normality, i.e $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^0) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta^0))$ (with p fixed).

Question: what happens when p is large relative to n ?

- If $p > n$, $\hat{\beta}^{\text{ols}}$ is not even defined.
- If $p \approx n$, $\text{Var}(\hat{\beta}^{\text{ols}})$ explodes since $X^T X$ is near singular.
- More generally, if $p, n \rightarrow \infty$ then asymptotic normality can break down.

Recall the bias-variance decomposition:

$$\begin{aligned} \text{mse}(\tilde{\beta}) &= \mathbb{E}_{\beta^0, \sigma^2} [(\tilde{\beta} - \beta^0)(\tilde{\beta} - \beta^0)] \\ &= \mathbb{E}_{\beta^0, \sigma^2} \left\| \tilde{\beta} - \mathbb{E}\tilde{\beta} + \mathbb{E}\tilde{\beta} - \beta^0 \right\|^2 \\ &= \text{Var}(\tilde{\beta}) + \left\| \mathbb{E}(\tilde{\beta}) - \beta^0 \right\|^2. \end{aligned}$$

We introduce bias to reduce the variance.

1.1 Ridge regression

Define

$$(\hat{\mu}_\lambda^R, \hat{\beta}_\lambda^R) = \operatorname{argmin}_{(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p} \left[\|Y - \mu \mathbf{1} - X\beta\|^2 + \underbrace{\lambda \|\beta\|^2}_{\text{penalty for large } \beta} \right].$$

λ is called a *regularisation* or *tuning* parameter. We shall assume the columns of X have been standardised (mean 0, variance 1).

After standardisation, we can show that

$$\hat{\mu}_\lambda^R = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}.$$

Hence, if we replace Y with $Y - \mathbf{1}\bar{Y}$ we can write

$$\begin{aligned} \hat{\beta}_\lambda^R &= \operatorname{argmin}_{\beta \in \mathbb{R}^p} [\|Y - X\beta\|^2 + \lambda \|\beta\|^2] \\ &= \underbrace{(X^T X + \lambda I_p)^{-1}}_{\text{always invertible}} X^T Y. \end{aligned}$$

Theorem 1.1. For $\lambda > 0$ sufficiently small,

$$\operatorname{mse}(\hat{\beta}^{ols}) - \operatorname{mse}(\hat{\beta}_\lambda^R) = \mathbb{E}\|\hat{\beta}^{ols} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}_\lambda^R - \beta^0\|^2 > 0. \quad (*)$$

Proof. We have

$$Y = X\beta^0 + \varepsilon.$$

The bias of $\hat{\beta}_\lambda^R$ is

$$\begin{aligned} \mathbb{E}(\hat{\beta}_\lambda^R - \beta^0) &= (X^T X + \lambda I)^{-1} X^T X \beta^0 - \beta^0 \\ &= (X^T X + \lambda I)^{-1} (X^T X + \lambda I - \lambda I) \beta^0 - \beta^0 \\ &= -\lambda (X^T X + \lambda I)^{-1} \beta^0. \end{aligned}$$

While we have variance

$$\begin{aligned} \operatorname{Var}(\hat{\beta}_\lambda^R) &= \mathbb{E} \|(X^T X + \lambda I)^{-1} X^T \varepsilon\|^2 \\ &= \sigma^2 [(X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}]. \end{aligned}$$

Then (*) becomes

$$\begin{aligned} &\mathbb{E}\|\hat{\beta}^{ols} - \beta^0\|^2 - \mathbb{E}\|\hat{\beta}_\lambda^R - \beta^0\|^2 \\ &= \sigma^2 (X^T X)^{-1} - \sigma^2 (X^T X + \lambda I) X^T X (X^T X + \lambda I)^{-1} \\ &\quad - \lambda^2 (X^T X + \lambda I)^{-1} \beta^0 (\beta^0)^T (X^T X + \lambda I)^{-1} \\ &= \vdots \quad \quad \quad (\text{use SVD } X = UDV^T) \\ &= \lambda (X^T X + \lambda I)^{-1} [\sigma^2 \{2I_p + \lambda (X^T X)^{-1}\} - \lambda \beta^0 (\beta^0)^T] (X^T X + \lambda I)^{-1}. \end{aligned}$$

We want to show this is positive definite. This is equivalent to

$$\begin{aligned}\sigma^2 [2I + \lambda(X^T X)^{-1}] - \lambda\beta^0(\beta^0)^T &> 0 \\ \iff 2\sigma^2 I - \lambda\beta^0(\beta^0)^T &> 0 \\ \iff 2\sigma^2 \|z\|^2 - \lambda(z^T \beta^0)^2 &> 0 \quad \forall z \in \mathbb{R}^p.\end{aligned}\tag{†}$$

We also have $(z^T \beta^0)^2 \leq \|z\|^2 \|\beta^0\|^2$ by Cauchy-Schwarz. Hence (†) holds for all $\lambda < \frac{2\sigma^2}{\|\beta^0\|^2}$. \square

Singular value decomposition

Suppose $n \geq p$, so we can always write $X \in \mathbb{R}^{n \times p}$ as

$$X = UDV^T \quad (\text{“thin SVD”})$$

where $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$, with orthonormal columns, $D \in \mathbb{R}^{p \times p}$ diagonal with $D_{11} \geq D_{22} \geq \dots \geq D_{pp} \geq 0$.

The fitted values in ridge regression are

$$\begin{aligned} \hat{Y}_\lambda^R &= X\hat{\beta}_\lambda^R = X(X^T X + \lambda I)^{-1} X^T Y \\ &= UDV^T (VD^2 V^T + \lambda I)^{-1} VDU^T Y \quad (\text{using } VV^T = V^T V = I) \\ &= UD(D^2 + \lambda I)^{-1} DU^T Y \\ &= \sum_{j=1}^p U_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} U_j^T Y \end{aligned}$$

where U_j denotes the j th column of U . For reference, in OLS regression

$$\hat{Y}^{ols} = X\hat{\beta}^{ols} = X(X^T X)^{-1} X^T Y = \sum_{j=1}^p U_j U_j^T Y.$$

So ridge “projects” onto columns of U , but it shrinks j th component by a factor

$$\frac{D_{jj}^2}{D_{jj}^2 + \lambda}.$$

Hence it shrinks small singular values to 0 rapidly.

Note. The matrix $X(X^T X)^{-1} X^T Y$ is known as the “hat matrix” and it represents an orthogonal projection onto the column space of X .

The SVD of X is related to principal component analysis.

Definition. The k th *principal component* $U^{(k)}$ of X and *principal direction* $v^{(k)}$ of X are defined recursively by

$$v^{(k)} = \operatorname{argmax}_{v \in \mathbb{R}^p} \|Xv\|^2 \text{ subject to } \|v\| = 1, (v^{(j)})^T X^T X v = 0 \quad \forall j < k$$

and

$$u^{(k)} = Xv^{(k)}.$$

Lemma 1.2. If $D_{jj} > 0$ for all $j \in \{1, \dots, p\}$ then $v^{(k)} = V_k$, $u^{(k)} = D_{kk} U_k$.

Message: ridge is good when the signal (β^0) is large for the top principal components of X .

Computation: we can compute \hat{Y}_λ^R for any value of λ quickly after doing an SVD, which has cost $\mathcal{O}(np^2)$.

1.2 v -fold cross-validation

We assume that (x_i, Y_i) , $i = 1, \dots, n$ is iid from some distribution (random design matrix). Let (x^*, Y^*) be another independent observation from this distribution. We may wish to pick λ minimising the mean-squared prediction error (MSPE) conditional on (X, Y) :

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2 | (X, Y)\}.$$

A less ambitious goal is to minimise the MSPE

$$\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2\} = \mathbb{E}\left[\mathbb{E}\{(Y^* - (x^*)^T \hat{\beta}_\lambda^R(X, Y))^2 | (X, Y)\}\right]. \quad (\ddagger)$$

We can try to estimate this quantity for different values of λ , using data splitting.

- Let $(X^{(1)}, Y^{(1)}), \dots, (X^{(v)}, Y^{(v)})$ be groups of data points of roughly equal size. These are called *folds*.
- Let $(X^{(-k)}, Y^{(-k)})$ denote all the folds except the k th.
- Let $\kappa(i)$ be the fold to which sample i (i.e. (X_i, Y_i)) belongs.

Our estimator of (\ddagger) is

$$\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i - x_i^T \underbrace{\hat{\beta}_\lambda^R(X^{(-\kappa(i))}, Y^{(-\kappa(i))})}_{\text{using all folds except the ones containing } (x_i, Y_i)} \right\}^2.$$

Then define

$$\lambda_{\text{CV}} = \operatorname{argmin}_{\lambda \in \{\lambda_1, \dots, \lambda_m\}} \text{CV}(\lambda).$$

We use the estimator

$$\hat{\beta}_{\lambda_{\text{CV}}}^R(X, Y).$$

How to choose v ?

Note.

- The expectation of each summand in $\text{CV}(\lambda)$ is almost the same as \ddagger , which is what we want to estimate. The only difference is the size of the training set. Hence the bias of $\text{CV}(\lambda)$ is small when v is large [the extreme of this is $v = n$, called “leave one out” cross-validation].
- When v is large, the estimator $\hat{\beta}_\lambda^R(X^{(-k)}, Y^{(-k)})$ is similar for different values of k , which leads to positively correlated summands in $\text{CV}(\lambda)$, leading to high variance.
- A common choice is $v = 5$ or $v = 10$.

1.3 Kernel trick

We have

$$\hat{Y}_\lambda^R = X(X^T X + \lambda I)^{-1} X^T Y.$$

Note that

$$\begin{aligned} X^T(XX^T + \lambda I) &= (X^T X + \lambda I)X^T \\ \implies (X^T X + \lambda I)^{-1} X^T &= X^T (XX^T + \lambda I)^{-1} \\ \implies X \underbrace{(X^T X + \lambda I)^{-1}}_{p \times p} X^T Y &= X X^T \underbrace{(XX^T + \lambda I)^{-1}}_{n \times n} Y. \end{aligned}$$

The computation cost of the LHS is $\mathcal{O}(np^2 + p^3)$ while the RHS is $\mathcal{O}(pn^2 + n^3)$.

- When $p \gg n$, the 2nd expression is cheaper to compute;
- The fitted values in ridge regression only depend on X through the “Gram matrix” $K = XX^T$, with entries $K_{ij} = \langle x_i, x_j \rangle$.

Suppose we wish to fit a quadratic model:

$$Y_i = x_i^T \beta + \sum_{k,l} x_{ik} x_{il} \theta_{kl} + \varepsilon_i.$$

This can be done with a linear model where we replace the predictors $x_i \in \mathbb{R}^p$ with a new “feature” vector:

$$\phi(x_i) = (x_{i1}, x_{i2}, \dots, x_{ip}, x_{i1}x_{i1}, x_{i1}x_{i2}, \dots, x_{ip}x_{ip}) \in \mathbb{R}^{p+p^2}.$$

We call ϕ a “feature map”. Now we have $\mathcal{O}(p^2)$ predictors. If $p^2 \gg n$, to compute ridge fitted values, we want to use the 2nd expression, with cost $\mathcal{O}(p^2 n^2 + n^3)$.

However, the part that scales as $\mathcal{O}(p^2 n^2)$ is just the computation of the Gram matrix with entries $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$.

The *kernel trick* offers a shortcut for computing K .

Idea:

$$\begin{aligned} \left(\frac{1}{2} + x_i^T x_j \right)^2 - \frac{1}{4} &= \left(\frac{1}{2} + \sum_k x_{ik} x_{jk} \right)^2 - \frac{1}{4} \\ &= \sum_k x_{ik} x_{jk} + \sum_{k,l} x_{ik} x_{il} x_{jk} x_{jl} \\ &= \langle \phi(x_i), \phi(x_j) \rangle = K_{ij}. \end{aligned}$$

The LHS can be computed in $\mathcal{O}(p)$ iterations, so we can obtain K in $\mathcal{O}(n^2 p)$ iterations, and we can compute the fitted values in ridge regression in $\mathcal{O}(n^2 p + n^3)$, which is not worse than the linear model!

Notes.

- For many feature maps ϕ , there are similar shortcuts.
- Instead of focusing on ϕ , we can directly think of the function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ as a measure of “similarity” between inputs x_i, x_j .

Question: for which similarities k is there a feature map ϕ such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle?$$

1.4 Kernels

Definition. An *inner product space* is a real vector space \mathcal{H} endowed with a map $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ satisfying:

- (i) Symmetry: for all $u, v \in \mathcal{H}$ we have $\langle u, v \rangle = \langle v, u \rangle$;
- (ii) Bilinearity: for all $a, b \in \mathbb{R}$ and all $u, v, w \in \mathcal{H}$ we have

$$\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle.$$

- (iii) Positive-definiteness: we have $\langle u, u \rangle \geq 0$ for all $u \in \mathcal{H}$, with equality if and only if $u = 0$.

Suppose that regression inputs x_1, \dots, x_n take values in an abstract set \mathcal{X} (so far we’ve had $\mathcal{X} = \mathbb{R}^p$, but the x_i ’s could be functions; images; graphs; etc.).

Goal: characterise similarity functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ for which there is an inner product space \mathcal{H} and a feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}.$$

Definition. A (*positive-definite*) *kernel* k is a symmetric map $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

Remark. A kernel is not an inner product on \mathcal{X} in general. Indeed, \mathcal{X} does not even need to be a vector space, and k need not be bilinear. However, we do have a version of the Cauchy-Schwarz inequality for kernels.

Proposition 1.3. *Let k be a kernel on \mathcal{X} . Then*

$$k(x, x')^2 \leq k(x, x)k(x', x') \quad \forall x, x' \in \mathcal{X}.$$

Proof. Since k is a kernel,

$$\begin{pmatrix} k(x, x) & k(x, x') \\ k(x', x) & k(x', x') \end{pmatrix} \geq 0.$$

Hence this has non-negative determinant and $k(x, x)k(x', x') - k(x, x')^2 \geq 0$. \square

Proposition 1.4. *Any similarity k defined by*

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad \forall x, x' \in \mathcal{X}$$

is a kernel.

Proof. Symmetry of k is clear. Let $x_1, \dots, x_n \in \mathcal{X}$ be arbitrary and take any vector $\alpha \in \mathbb{R}^n$. We need to show $\alpha^T K \alpha \geq 0$. Indeed

$$\begin{aligned} \alpha^T K \alpha &= \sum_{i,j} \alpha_i K_{ij} \alpha_j \\ &= \sum_{i,j} \alpha_i \langle \phi(x_i), \phi(x_j) \rangle \alpha_j \\ &= \left\langle \sum_{i=1}^n \alpha_i \phi(x_i), \sum_{j=1}^n \alpha_j \phi(x_j) \right\rangle && \text{(linearity of } \langle \cdot, \cdot \rangle \text{)} \\ &\geq 0. && \text{(positive-definiteness of } \langle \cdot, \cdot \rangle \text{)} \end{aligned}$$

□

Examples of kernels

Proposition 1.5 (Closure property). *Suppose k_1, k_2, \dots are kernels on \mathcal{X} . Then*

- (i) *If $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel. If $k(x, x') := \lim_{m \rightarrow \infty} k_m(x, x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel.*
- (ii) *The pointwise product $k(x, x') = k_1(x, x')k_2(x, x')$ is a kernel.*

Proof. Example Sheet 1. □

Some examples of kernels are:

- Linear kernel: $k(x, x') = x^T x'$ (for $\mathcal{X} = \mathbb{R}^p$);
- Polynomial kernel: $k(x, x') = (1 + x^T x')^d$, $d \in \mathbb{N}$ ($\mathcal{X} = \mathbb{R}^p$). Note $(x, x') \mapsto 1$ is a kernel so this is a kernel by the previous proposition;
- Gaussian kernel: $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$, $\sigma^2 > 0$ the *bandwidth* of the kernel. Indeed note

$$\exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \underbrace{\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)}_{:=k_1(x, x')} \underbrace{\exp\left(-\frac{\|x'\|^2}{2\sigma^2}\right)}_{:=k_2(x, x')} \exp\left(\frac{x^T x'}{\sigma^2}\right).$$

It suffices to show k_1, k_2 are kernels. For k_1 we have $k_1(x, x') = \langle \phi(x), \phi(x') \rangle$ where $\phi : \mathbb{R}^p \rightarrow \mathbb{R}$ is defined by

$$\phi(x) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right).$$

For k_2 we have that $(x, x') \mapsto x^T x'$ is a kernel and k_2 can be Taylor expanded so is the limit of kernels;

- Sobolev kernel: let $\mathcal{X} = [0, 1]$ and set $k(x, x') = \min(x, x') = \text{Cov}(Wx, Wx')$ where $(W_t)_{t \geq 0}$ a Brownian motion (positive definite as a covariance);
- Jaccard similarity kernel: let $\mathcal{X} = \mathcal{P}(\{1, \dots, p\})$ and set

$$k(x, x') = \begin{cases} \frac{|x \cap x'|}{|x \cup x'|} & \text{if } x \cup x' \neq \emptyset \\ 0 & \text{otherwise} \end{cases}.$$

(For proof this is a kernel see Example Sheet 1.)

Remark. There is no finite-dimensional feature map $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^m$ representing the Gaussian kernel.

Theorem 1.6 (Moore-Aronzajn Theorem). *For every kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, there exists a feature map ϕ taking values in some inner product space \mathcal{H} such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathcal{X}$.*

Proof. Take \mathcal{H} to be the vector space of functions from \mathcal{X} to \mathbb{R} of the form

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}.$$

In other words, \mathcal{H} is the linear span of functions of the form $f(\cdot, x)$ for $x \in \mathcal{X}$. Our feature map $\phi : \mathcal{X} \rightarrow \mathcal{H}$ will be $\phi(x) = k(\cdot, x)$. We now define the inner product $\langle \cdot, \cdot \rangle$ on \mathcal{H} . Let

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}$$

and

$$g(\cdot) = \sum_{j=1}^m \beta_j k(\cdot, x'_j).$$

Then define

$$\langle f, g \rangle := \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j) = \sum_{i=1}^n \alpha_i g(x_i) = \sum_{j=1}^m \beta_j f(x'_j).$$

In particular, the final two expressions show $\langle \cdot, \cdot \rangle$ is well-defined (it doesn't matter how we represent f, g as these linear combinations).

We observe directly from the definition that $\langle \phi(x), \phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$ as required. We must show $\langle \cdot, \cdot \rangle$ is indeed an inner product. It is certainly bilinear and symmetric. So we show it is positive-definite. Note that

$$\langle f, f \rangle = \sum_{i=1}^n \sum_{j=1}^n \alpha_i k(x_i, x_j) \alpha_j \geq 0 \quad (\dagger)$$

since k is a kernel. It remains to show $\langle f, f \rangle$ implies $f(x) = 0$ for all $x \in \mathcal{X}$.

Note that $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ is a kernel. Indeed, given functions f_1, \dots, f_m and $\gamma_1, \dots, \gamma_n \in \mathbb{R}$ we have

$$\sum_{i=1}^n \sum_{j=1}^n \gamma_i \langle f_i, f_j \rangle \gamma_j = \left\langle \sum_{i=1}^n \gamma_i f_i, \sum_{j=1}^n \gamma_j f_j \right\rangle \geq 0$$

by (\dagger) .

Now note that

$$f(x)^2 = \langle f, k(\cdot, x) \rangle^2 \leq \langle f, f \rangle \langle k(\cdot, x), k(\cdot, x) \rangle$$

by the Cauchy-Schwarz property for kernels. Hence $\langle f, f \rangle = 0$ implies $f(x) = 0$ for all $x \in \mathcal{X}$. \square

Remark. The space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ constructed in the proof has the property that

$$f(x) = \langle f, \underbrace{k(\cdot, x)}_{\phi(x)} \rangle.$$

As a consequence

$$|f(x) - g(x)| = |\langle f - g, k(\cdot, x) \rangle| \leq \|f - g\|_{\mathcal{H}} k(x, x)^{1/2}.$$

Hence convergence in $(\mathcal{H}, \|\cdot\|)$ implies pointwise convergence.

Lemma 1.7. Let \mathcal{H} be a Hilbert space and $\mathcal{V} \subseteq \mathcal{H}$ a closed subspace. Then $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp$, i.e for any $f \in \mathcal{H}$ we have $f = u + v$ where $u \in \mathcal{V}$ and $v \in \mathcal{V}^\perp$ and u, v are unique.

Proof. See Part II Linear Analysis. \square

Definition. A Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ is a *reproducing kernel Hilbert space* (RKHS) if for all $x \in \mathcal{X}$, there exists $k_x \in \mathcal{H}$ such that $f(x) = \langle k_x, f \rangle$ for all $f \in \mathcal{H}$.

The function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by $(x, x') \mapsto \langle k_x, k_{x'} \rangle = k_x(x')$ is known as the reproducing kernel of \mathcal{H} .

Remark. By the Riesz Representation Theorem, it is equivalent to define an RKHS as a Hilbert space where the evaluation operator $E_x : f \mapsto f(x)$ is a continuous linear operator.

The Moore-Aronzajn Theorem says that whenever k is a kernel, there is an inner product space \mathcal{H} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ where $f(x) = \langle f, k(\cdot, x) \rangle$ and thus $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle$.

This implies that if $(f_n)_{n \geq 1}$ is Cauchy in \mathcal{H} ,

$$|f_n(x) - f_m(x)| \leq \sqrt{k(x, x)} \|f_n - f_m\|_{\mathcal{H}} \rightarrow 0.$$

Hence $(f_n)_{n \geq 1}$ has a pointwise limit $f^* : \mathcal{X} \rightarrow \mathbb{R}$ by completeness of \mathbb{R} . So we can complete \mathcal{H} by including all limits of Cauchy sequences (Hausdorff completion) to obtain a Hilbert space $\overline{\mathcal{H}}$. By construction, $\overline{\mathcal{H}}$ is a RKHS with reproducing kernel k .

Proposition 1.8. If \mathcal{G} is a RKHS of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $\mathcal{G} \supseteq \mathcal{H}$, then $\overline{\mathcal{H}} = \mathcal{G}$.

Proof. Example Sheet 1. \square

Notation: from now on the RKHS is \mathcal{H} (i.e $\mathcal{H} = \overline{\mathcal{H}}$).

Examples.

- Linear kernel: $k(x, x') = x^T x'$. Then $\mathcal{H} = \{f : f(x) = x^T \beta, \beta \in \mathbb{R}^p\}$. If $f(x) = x^T \beta$ then $\|f\|_{\mathcal{H}}^2 = \|\beta\|^2$.
- Sobolev kernel: $k(x, x') = \min(x, x')$ with $\mathcal{X} = [0, 1]$. Then \mathcal{H} is the space of continuous functions $f : [0, 1] \rightarrow \mathbb{R}$ with $f(0) = 0$, for which

$$\int_0^1 |f'(x)|^2 dx < \infty$$

where f' is the weak derivative.

The Representer Theorem

If \mathcal{H} is the RKHS of the linear kernel, we can express ridge regression as

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \left\{ \sum_{i=1}^n \underbrace{(Y_i - f(x_i))^2}_{x_i^T \beta} + \lambda \underbrace{\|f\|_{\mathcal{H}}^2}_{\|\beta\|^2} \right\}.$$

In *kernel ridge regression*, we solve this problem in a more general RKHS with kernel k , e.g the Gaussian kernel.

Theorem 1.9 (Representer Theorem). *Let:*

- $c : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be an arbitrary loss;
- $J : [0, \infty) \rightarrow \mathbb{R}$ be strictly increasing;
- $x_1, \dots, x_n \in \mathcal{X}$, $Y \in \mathbb{R}^n$;
- \mathcal{H} an RKHS with representing kernel k ;
- $K_{ij} = k(x_i, x_j)$, $i, j \in [n]$.

Then \hat{f} minimises

$$Q_1(f) = c(Y, x_1, \dots, x_n, f(x_1), \dots, f(x_n)) + J(\|f\|_{\mathcal{H}}^2)$$

over $f \in \mathcal{H}$ if and only if $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ and $\hat{\alpha}$ minimises Q_2 over $\alpha \in \mathbb{R}^n$ where

$$Q_2(\alpha) = c(Y, x_1, \dots, x_n, K\alpha) + J(\alpha^T K \alpha).$$

Example. In kernel ridge regression we just need to solve the quadratic program

$$\hat{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}^n} \|Y - K\alpha\|^2 + \lambda \alpha^T K \alpha = (K + I\lambda)^{-1}.$$

Then the fitted values are given by $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$.

Intuition: to make a prediction at “test” point x^* the terms in $\hat{f}(x^*)$ that contribute the most are those for training points x_i with similarity $k(x^*, x_i)$ large.

Proof of the Representer Theorem. Note $V = \operatorname{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\}$ is a closed (as its finite dimensional) subspace of \mathcal{H} . Hence any $f \in \mathcal{H}$ can be written as $f = u + v$ for $u \in V$ and $v \in V^\perp$.

We have $f(x_i) = \langle k(\cdot, x_i), u + v \rangle = \langle k(\cdot, x_i), u \rangle = u(x_i)$. Then

$$\|f\|_{\mathcal{H}}^2 = \|v\|_{\mathcal{H}}^2 + \|u\|_{\mathcal{H}}^2.$$

In the expression for Q_1 , the first term only depends on u , and the second term is $J(\|f\|_{\mathcal{H}}^2) \geq J(\|u\|_{\mathcal{H}}^2)$ with equality if and only if $v = 0$. Hence any minimiser

of Q_1 is contained in \mathcal{V} .

So write $f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for the minimiser. Now note

$$(f(x_1), \dots, f(x_n)) = K\alpha$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i k(x_i, x_j) \alpha_j = \alpha^T K \alpha$$

so therefore for any $f \in \mathcal{V}$, $Q_1(f) = Q_2(\alpha)$. Hence $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ minimises Q_1 if and only if $\hat{\alpha}$ minimises Q_2 . \square

Now we will assume that

$$Y_i = f^0(x_i) + \varepsilon_i, \quad \mathbb{E}\varepsilon = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I$$

where $\|f^0\|_{\mathcal{H}} \leq 1$.

Note. This is equivalent to $cY_i = cf^0(x_i) + c\varepsilon_i$ so $\|cf^0\|_{\mathcal{H}} = c\|f^0\|_{\mathcal{H}}$, $\text{Var}(c\varepsilon_i) = \sigma^2 c^2$. So the “signal-to-noise ratio” is

$$\frac{\text{Var}(c\varepsilon_i)}{\|cf^0\|_{\mathcal{H}}^2} = \frac{\text{Var}(\varepsilon_i)}{\|f^0\|_{\mathcal{H}}^2} \geq \sigma^2.$$

Theorem 1.10. *Let K have eigenvalues $d_1 \geq d_2 \geq \dots \geq d_n \geq 0$. Then*

$$\begin{aligned} \text{MSPE}(\hat{f}_n) &= \frac{1}{n} \mathbb{E} \left\{ \sum_{i=1}^n (f^0(x_i) - \hat{f}_n(x_i))^2 \right\} \\ &\leq \frac{\sigma^2}{n} \sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda}{4n} \\ &\leq \frac{\sigma^2}{n} \frac{1}{\lambda} \sum_{i=1}^n \min\left(\frac{d_i}{4}, \lambda\right) + \frac{\lambda}{4n}. \end{aligned}$$

Proof. From the Representer Theorem $(\hat{f}_n(x_1), \dots, \hat{f}_n(x_n))^T = K(K + \lambda I)^{-1}Y$. As $f^0 \in \mathcal{H}$ we have $(f^0(x_1), \dots, f^0(x_n))^T = K\alpha$ for some $\alpha \in \mathbb{R}^n$ (see Example Sheet). Moreover, $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha$. Let the UDU^T be the eigen-decomposition of K , with $D_{ii} = d_i$. Define $\Theta = U^T K \alpha$. Then

$$\begin{aligned} n\text{MSPE}(\hat{f}_n) &= \mathbb{E} \left\| K(K + \lambda I)^{-1} \underbrace{(U\Theta + \varepsilon)}_Y - \underbrace{U\Theta}_{(f^0(x_1), \dots, f^0(x_n))^T} \right\|^2 \\ &= \mathbb{E} \|UDU^T(UDU^T + \lambda I)^{-1}(U\Theta + \varepsilon) - U\Theta\|^2 \\ &= \mathbb{E} \|D(D + \lambda I)^{-1}(\Theta + U^T \varepsilon) - \Theta\|^2 \quad (U^T U = I) \\ &= \underbrace{\mathbb{E} \|\{D(D + \lambda I)^{-1} - I\}\varepsilon\|^2}_{:= (1)} + \underbrace{\mathbb{E} \|D(D + \lambda I)^{-1}U^T \varepsilon\|^2}_{:= (2)}. \quad (\mathbb{E}\varepsilon = 0) \end{aligned}$$

So

$$\begin{aligned} (2) &= \mathbb{E} [\{D(D + \lambda I)^{-1}U^T \varepsilon\}^T \{D(D + \lambda I)^{-1}U^T \varepsilon\}] \\ &= \mathbb{E} [\text{tr}(\{D(D + \lambda I)^{-1}U^T \varepsilon\}^T \{D(D + \lambda I)^{-1}U^T \varepsilon\})] \\ &= \mathbb{E} [\text{tr}(D(D + \lambda I)^{-1} \varepsilon \varepsilon^T D(D + \lambda I)^{-1})] \quad (\text{circular property of tr}) \\ &= \text{tr}(D(D + \lambda I)^{-1} \sigma^2 I D(D + \lambda I)^{-1}) \\ &= \sigma^2 \sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2}. \end{aligned}$$

Also

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

Since $\Theta = DU^T \alpha$, so if $d_i = 0$ then $\Theta_i = 0$. So let D^+ be a diagonal matrix with $D_{ii}^+ = \begin{cases} d_i^{-1} & \text{if } d_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$.

Then,

$$\begin{aligned} \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} &= \|\sqrt{D^+} \Theta\|^2 = \alpha^T K U D^+ U^T K \alpha \\ &= \alpha^T U D D^+ D U^T \alpha \\ &= \alpha^T U D U^T \alpha \quad (D D^+ D = D) \\ &= \alpha^T K \alpha \leq 1. \end{aligned}$$

Then

$$\begin{aligned} (1) &= \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \leq \max_{1 \leq i \leq n} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \sum_{i:d_i > 0} \frac{\Theta_i^2}{d_i} \\ &\leq \max_{1 \leq i \leq n} \frac{d_i \lambda^2}{(d_i + \lambda)^2} \\ &\leq \frac{\lambda}{4}. \quad ((a+b)^2 \geq 4ab) \end{aligned}$$

Combining the bounds for (1) and (2) gives the first inequality. Finally, for the final inequality we note that

$$\frac{d_i^2}{(d_i + \lambda)^2} \leq \min \left\{ 1, \frac{d_i^2}{4d_i \lambda} \right\} = \frac{1}{\lambda} \min \left\{ \lambda, \frac{d_i}{4} \right\}.$$

□

Question: when is the upper bound good?

Random design

Let $(\mathcal{X}, \mathcal{B}, \mathbb{P})$ be a probability space, where \mathcal{X} is a metric space, \mathcal{B} is the Borel σ -algebra on \mathcal{X} . Assume that $x_1, \dots, x_n \sim^{\text{iid}} \mathbb{P}$.

Theorem 1.11 (Mercer's Theorem). *Under mild assumptions on k, \mathbb{P} , there is an orthonormal basis (e_i) of $\mathcal{L}^2(\mathbb{P})$, i.e*

$$\int_{\mathcal{X}} e_l(x) e_j(x) d\mathbb{P}(x) = \mathbb{1}\{l = j\}$$

and eigenvalues (μ_i) with $\sum_{i=1}^n \mu_i < \infty$ such that

$$\mu_j e_j(x') = \int_{\mathcal{X}} k(x, x') e_j(x) d\mathbb{P}(x).$$

Furthermore

$$k(x, x') = \sum_{l=1}^{\infty} \mu_l e_l(x) e_l(x')$$

and this series is absolutely convergent.

Proof. Not given. □

Let $\hat{\mu}_1, \dots, \hat{\mu}_n$ be (random) eigenvalues of K/n . As it turns out, when n is large $\hat{\mu}_i \approx \mu_i$. Let $\gamma = \lambda/n$, then a previous theorem gives

$$\text{MSPE}(\hat{f}_{\gamma n}) \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^n \min\left(\frac{\hat{\mu}_i}{4}, \gamma\right) + \frac{\gamma}{4}.$$

Then the MSPE is a random variable depending on x_1, \dots, x_n .

Lemma 1.12.

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

Proof. Not given. □

This lemma means we can bound

$$\underbrace{\mathbb{E}[\text{MSPE}(\hat{f}_{\gamma n})]}_{\text{over } Y \text{ and } x_1, \dots, x_n} \leq \frac{\sigma^2}{\gamma} \frac{1}{n} \sum_{i=1}^n \min\left(\frac{\mu_i}{4}, \gamma\right) + \frac{\gamma}{4}.$$

Theorem 1.13. *Under the assumptions of Mercer's Theorem, there is a sequence $(\gamma_n)_{n \geq 1}$ such that for fixed $\sigma^2 > 0$,*

$$\frac{1}{n} \mathbb{E} \left\{ \sum_{i=1}^n (f^0(x_i) - \hat{f}_{\gamma n}(x_i))^2 \right\} = o(n^{-1/2}) \text{ as } n \rightarrow \infty.$$