1. Consider minimising the following objective involving response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$ over $(\mu, \beta) \in \mathbb{R} \times \mathbb{R}^p$:
$$\|Y - \mu\mathbf{1} - X\beta\|_2^2 + J(\beta).$$
Here $J : \mathbb{R}^p \to \mathbb{R}$ is an arbitrary penalty function. Suppose $\bar{X}_k = 0$ for $k = 1, \ldots, p$. Assuming that a minimiser $(\hat{\mu}, \hat{\beta})$ exists, show that $\hat{\mu} = \bar{Y}$. Now take $J(\beta) = \lambda\|\beta\|_2^2$ so we have the ridge regression objective. Show that
$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y.$$
From here onwards, whenever we refer to ridge regression, we will assume X has had its columns mean-centred.

2. Consider performing ridge regression when $Y = X\beta^0 + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ has full column rank, and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Let the SVD of X be UDV^T and write $U^T X\beta^0 = \gamma$. Show that
$$\frac{1}{n}\mathbb{E}\|X\beta^0 - X\hat{\beta}_\lambda^R\|_2^2 = \frac{1}{n}\sum_{j=1}^p \left(\frac{\lambda}{\lambda + D_{jj}^2}\right)^2 \gamma_j^2 + \frac{\sigma^2}{n}\sum_{j=1}^p \frac{D_{jj}^4}{(\lambda + D_{jj}^2)^2}.$$
Now suppose the size of the signal is n, so $\|X\beta^0\|_2^2 = n$. For what γ is the mean squared prediction error above minimised? For what γ is it maximised?

3. Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set with $\sqrt{\lambda}I$ added to the bottom of X (where I here is $p \times p$), and p zeroes added to the end of the response Y.

4. In the following, assume that forming AB where $A \in \mathbb{R}^{a\times b}$, $B \in \mathbb{R}^{b\times c}$ requires $O(abc)$ computational operations, and that if $M \in \mathbb{R}^{d\times d}$ is invertible, then forming M^{-1} requires $O(d^3)$ operations.
   (a) Suppose we wish to apply ridge regression to data $(Y, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$ with $n \gg p$. A complication is that the data is split into m separate datasets of size $n/m \in \mathbb{N}$,
$$Y = \begin{pmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{pmatrix} \qquad X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(m)} \end{pmatrix},$$
   with each dataset located on a different server. Moving large amounts of data between servers is expensive. Explain how one can produce ridge estimates $\hat{\beta}_\lambda$ by communicating only $O(p^2)$ numbers from each server to some central server. What is the total order of the computation time required at each server, and at the central server for your approach?
   (b) Now suppose instead that $p \gg n$ and it is instead the variables that are split across m servers, so each server has only a subset of $p/m \in \mathbb{N}$ variables for each observation, and some central server stores Y. Explain how one can obtain the fitted values $X\hat{\beta}_\lambda$ communicating only $O(n^2)$ numbers from each server to the central server. What is the total order of the computation time required at each server, and at the central server for your approach?

5. Prove Proposition 4 in our notes. *Hint: For part (ii) it may help to consider the eigendecompositions of positive semi-definite matrices* $K^{(1)}$ *and* $K^{(2)}$ *derived from kernels* k_1 *and* k_2 *in the form* $K^{(1)} = PDP^T = \sum_{i=1}^n P_i P_i^T D_{ii}$ *for example.*

6. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 < 1\}$. Show that $k(x, x') = (1 - x^T x')^{-\alpha}$ defined on $\mathcal{X} \times \mathcal{X}$, where $\alpha > 0$, is a kernel.

7. Suppose we have a matrix of predictors $X \in \mathbb{R}^{n\times p}$ where $p \gg n$. Explain how to obtain the fitted values of the following ridge regression using the kernel trick:

$$\text{Minimise over } \beta \in \mathbb{R}^p, \theta \in \mathbb{R}^{p(p-1)/2}, \gamma \in \mathbb{R}^p,$$
$$\sum_{i=1}^n \left(Y_i - \sum_{k=1}^p X_{ik}\beta_k - \sum_{k=1}^p \sum_{j=1}^{k-1} X_{ik}X_{ij}\theta_{jk} - \sum_{k=1}^p X_{ik}^2 \gamma_k \right)^2 + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\theta\|_2^2 + \lambda_3 \|\gamma\|_2^2.$$

Note we have indexed θ with two numbers for convenience.

8. Let $\hat{\alpha}$ be a minimiser of $\|Y - K\alpha\|_2^2 + \lambda\alpha^T K\alpha$ over α, with K being a kernel matrix as usual (i.e. symmetric positive semi-definite). Show that $K\hat{\alpha} = K(K+\lambda)^{-1}Y$.

9. Consider minimising
$$c(Y, X, f(x_1) + \mu, \ldots, f(x_n) + \mu) + J(\|f\|_{\mathcal{H}}^2)$$
over $f \in \mathcal{H}$ and $\mu \in \mathbb{R}$ where $\mathcal{H}$ is an RKHS. Here c is an arbitrary loss function and J is strictly increasing. Let k be the reproducing kernel of $\mathcal{H}$. Show that any minimiser $\hat{g}(\cdot) = \hat{f}(\cdot) + \hat{\mu}$ may be written as
$$\hat{g}(\cdot) = \hat{\mu} + \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$$
where $\hat{\alpha}_i \in \mathbb{R}$ for $i = 1, \ldots, n$.

10. This question proves a result needed for Theorem 7 in our notes. Let $\mathcal{H}$ be a RKHS of functions on $\mathcal{X}$ with reproducing kernel k and suppose $f^0 \in \mathcal{H}$. Let $x_1, \ldots, x_n \in \mathcal{X}$ and let K be the kernel matrix $K_{ij} = k(x_i, x_j)$. Show that
$$\left(f^0(x_1), \ldots, f^0(x_n)\right)^T = K\alpha,$$
for some $\alpha \in \mathbb{R}^n$ and moreover that $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K\alpha$.

11. Show from first principles that the Sobolev kernel is indeed a (positive definite) kernel.

12. Let $\mathcal{H}$ be an RKHS with reproducing kernel k. Show that if $h_x \in \mathcal{H}$ has the property that $\langle h_x, f\rangle = f(x)$ for all $f \in \mathcal{H}$, then $h_x(\cdot) = k(\cdot, x)$.

13. Prove that if k is a reproducing kernel for RKHS's $\mathcal{H}_1$ and $\mathcal{H}_2$, then $\mathcal{H}_1 = \mathcal{H}_2$, so the RKHS is uniquely determined by k. *Hint: First argue that it is enough to show the result for $\mathcal{H}_1 \subseteq \mathcal{H}_2$. Next consider decomposing each $f \in \mathcal{H}_2$ as $f = u + v$ with $u \in \mathcal{H}_1$ and $v \in \mathcal{H}_1^\perp$ and argue that $v = 0$.*