1. 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 $\|f^0\|_{\mathcal{H}}^2 \geq \alpha^T K \alpha$.
Solution: Let $V = \text{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$ and write $f^0 = u + v$ where $u \in V$ and $v \in V^\perp$. Then
$$f^0(x_i) = \langle f^0, k(\cdot, x_i) \rangle = \langle u, k(\cdot, x_i) \rangle.$$
Write $u = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. Then

$$f^0(x_i) = \sum_{j=1}^n \alpha_j \langle k(\cdot, x_j), k(\cdot, x_i) \rangle = \sum_{j=1}^n \alpha_j k(x_j, x_i) = K_i^T \alpha,$$

where K_i is the ith column (or row) of K. Thus $K\alpha = \left(f^0(x_1), \ldots, f^0(x_n) \right)^T$. By Pythagoras' theorem

$$\|f^0\|_{\mathcal{H}}^2 = \|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2 \geq \|u\|_{\mathcal{H}}^2 = \alpha^T K \alpha.$$

2. Let $Y \in \mathbb{R}^n$ be a vector of responses, $\Phi \in \mathbb{R}^{n \times d}$ a design matrix, $J : [0, \infty) \to [0, \infty)$ a strictly increasing function and $c : \mathbb{R}^n \times \mathbb{R}^n$ some cost function. Set $K = \Phi\Phi^T$. Show, without using the representer theorem, that $\hat{\theta}$ minimises

$$Q_1(\theta) := c(Y, \Phi\theta) + J(\|\theta\|_2^2)$$

over $\theta \in \mathbb{R}^p$ if and only if $\Phi\hat{\theta} = K\hat{\alpha}$ and $\hat{\alpha}$ minimises

$$Q_2(\alpha) := c(Y, K\alpha) + J(\alpha^T K \alpha)$$

over $\alpha \in \mathbb{R}^n$.
Solution: Suppose $\hat{\theta}$ minimises Q_1. Let $\Pi \in \mathbb{R}^{d \times d}$ be the orthogonal projection on to the row space of Φ. Then $\Phi\hat{\theta} = \Phi\Pi\hat{\theta}$ but $\|\hat{\theta}\|_2^2 = \|\Pi\hat{\theta}\|_2^2 + \|(I - \Pi)\hat{\theta}\|_2^2$. As J is strictly increasing, and by minimality of $\hat{\theta}$, we must have $(I - \Pi)\hat{\theta} = 0$. Thus $\hat{\theta}$ lies in the row space of Φ, so $\hat{\theta} = \Phi^T\alpha$, for some $\alpha \in \mathbb{R}^n$. We see then that

$$Q_1(\hat{\theta}) = c(Y, K\alpha) + J(\alpha^T K \alpha) = Q_2(\alpha).$$

Thus by minimality of $\hat{\theta}$, α must be such that it minimises Q_2. Now suppose $\hat{\alpha}$ minimises Q_2. Write $\hat{\theta} = \Phi^T\hat{\alpha}$, and note that $Q_1(\hat{\theta}) = Q_2(\hat{\alpha})$. Suppose $\tilde{\theta}$ has $Q_1(\tilde{\theta}) \leq Q_1(\hat{\theta})$. Then from the argument above, $Q_1(\Pi\tilde{\theta}) \leq Q_1(\tilde{\theta})$, and we may write $\Pi\tilde{\theta} = \Phi^T\tilde{\alpha}$ for some $\tilde{\alpha} \in \mathbb{R}^n$. But then we have

$$Q_2(\hat{\alpha}) = Q_1(\hat{\theta}) \geq Q_1(\tilde{\theta}) \geq Q_1(\Pi\tilde{\theta}) = Q_2(\tilde{\alpha}) \geq Q_2(\hat{\alpha}),$$

the last inequality following from minimality of $\hat{\alpha}$.

3. Let $x, x' \in \mathbb{R}^p$ and let $\psi \in \{-1, 1\}^p$ be a random vector with independent components taking the values $-1, 1$ each with probability $1/2$. Show that $\mathbb{E}(\psi^T x \psi^T x') = x^T x'$. Construct a random feature map $\hat{\phi} : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\{\hat{\phi}(x)\hat{\phi}(x')\} = (x^T x')^2$.
   Solution: $\mathbb{E}(x^T \psi \psi^T x') = x^T \mathbb{E}(\psi \psi^T) x' = x^T x'$. Given two vectors independent vectors $\psi^{(1)}, \psi^{(2)}$ each with the same distribution as ψ, set $\hat{\phi}(x) = x^T \psi^{(1)} x^T \psi^{(2)}$.

4. Let $\mathcal{X}$ be the set of all subsets of $\{1, \ldots, p\}$ and let $z, z' \in \mathcal{X}$. Let k be the Jaccard similarity kernel. Let π be a random permutation of $\{1, \ldots, p\}$. Let $M = \min\{\pi(j) : j \in z\}$, $M' = \min\{\pi(j) : j \in z'\}$. Show that

$$\mathbb{P}(M = M') = k(z, z'),$$

   when $z, z' \neq \emptyset$. Now let $\psi \in \{-1, 1\}^p$ be a random vector with i.i.d. components taking the values -1 or 1, each with probability $1/2$. By considering $\mathbb{E}(\psi_M \psi_{M'})$ show that the Jaccard similarity kernel is indeed a kernel. Explain how we can use the ideas above to approximate kernel ridge regression with Jaccard similarity, when n is very large (you may assume that none of the data points are the empty set).
   Solution: Let $H = \underset{k \in z \cup z'}{\operatorname{argmin}} \pi(k)$. Then

$$\mathbb{P}(M = M') = \mathbb{P}(M = M' = \pi(H)) = \mathbb{P}(H \in z \cap z') = \frac{|z \cap z'|}{|z \cup z'|}.$$

   Now

$$\mathbb{E}(\psi_M \psi_{M'}) = \mathbb{E}(\psi_M \psi_{M'} | M = M')\mathbb{P}(M = M') + \mathbb{E}(\psi_M \psi_{M'} | M \neq M')\mathbb{P}(M \neq M') = \mathbb{P}(M = M').$$

   Given $z_1, \ldots, z_n \in \mathcal{X} \setminus \{\emptyset\}$, let $M_i = \min\{\pi(k) : k \in z_i\}$ and define $S \in \{-1, 1\}^n$ by $S_i = \psi_{M_i}$. Let $K \in \mathbb{R}^{n \times n}$ have $K_{ij} = k(z_i, z_j)$. Then from the above we know $K = \mathbb{E}(SS^T)$ showing that it is positive semi-definite. If we have $z'_1, \ldots, z'_m = \emptyset$, the kernel matrix corresponding to $z_1, \ldots, z_n, z'_1, \ldots, z'_m$ would be block diagonal with one block corresponding to K above, and the other block being a matrix of ones. As both blocks are positive-semidefinite, k must be a kernel.

5. Consider the logistic regression model where we assume $Y_1, \ldots, Y_n \in \{-1, 1\}$ are independent and

$$\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^T \beta^0.$$

   Show that the maximum likelihood estimate $\hat{\beta}$ minimises

$$\sum_{i=1}^n \log\{1 + \exp(-Y_i x_i^T \beta)\}$$

   over $\beta \in \mathbb{R}^p$.
   Solution: Fix i and let $u = x_i^T \beta^0$. We have

$$\mathbb{P}(Y_i = 1) = \frac{e^u}{1 + e^u} = \frac{1}{1 + e^{-u}},$$

   and $\mathbb{P}(Y_i = -1) = 1/(1 + e^u)$, so $\mathbb{P}(Y_i = y) = 1/(1 + e^{-uy})$, from which the result easily follows by noting that the Y_i are independent.

6. Show that if W is mean-zero and sub-Gaussian with parameter σ, then $\text{Var}(W) \leq \sigma^2$.
   Solution: We know that $\mathbb{E}(e^{\alpha W}) \leq e^{\alpha^2 \sigma^2/2}$ for all $\alpha \in \mathbb{R}$. A power series expansion both sides of the above inequality yields

$$1 + \alpha \mathbb{E}(W) + \alpha^2 \mathbb{E}(W^2)/2 + \alpha^2 \mathbb{E}\left(\sum_{r=3}^{\infty} \frac{\alpha^{r-2} W^r}{r!} \right) \leq 1 + \alpha^2 \sigma^2/2 + O(\alpha^4).$$

Subtracting 1 and dividing by α^2 gives

$$\mathbb{E}(W^2)/2 + \mathbb{E}\left(\sum_{r=2}^{\infty} \frac{\alpha^{r-2} W^r}{r!} \right) \leq \sigma^2/2 + O(\alpha^2).$$

Now $e^W + e^{-W} \geq e^{|W|}$ so $e^{|W|}$ is integrable and $e^{|W|} \geq \sum_{r=3}^{\infty} \frac{\alpha^{r-2} W^r}{r!}$ for all $\alpha < 1$. Thus by dominated convergence theorem, $\mathbb{E}(\sum_{r=3}^{\infty} \frac{\alpha^{r-2} W^r}{r!}) \to 0$ as $\alpha \to 0$.

7. Verify Hoeffding's lemma for the special case where W is a Rademacher random variable, so W takes the values $-1, 1$ each with probability $1/2$.
   Solution:

$$\mathbb{E} e^{\alpha W} = \frac{1}{2}(e^{-\alpha} + e^{\alpha}) = \sum_{r=0}^{\infty} \frac{\alpha^{2r}}{(2r)!} \leq \sum_{r=0}^{\infty} \frac{\alpha^{2r}}{2^r r!} = e^{\alpha^2/2}.$$

8. (a) Let $W \sim \chi_d^2$. Show that

$$\mathbb{P}(|W/d - 1| \geq t) \leq 2 e^{-dt^2/8}$$

for $t \in (0, 1)$. You may use the facts that the mgf of a χ_1^2 random variable is $1/\sqrt{1 - 2\alpha}$ for $\alpha < 1/2$, and $e^{-\alpha}/\sqrt{1 - 2\alpha} \leq e^{2\alpha^2}$ when $|\alpha| < 1/4$.
   Solution: Note that W is the sum of d independent χ_1^2-distributed random variables. Thus

$$\mathbb{E} e^{\alpha(W-d)} = (1 - 2\alpha)^{-d/2} e^{-\alpha d}$$

for $\alpha < 1/2$. Suppose $t \in (0, 1)$. Using the Chernoff bound, we get

$$\begin{aligned}
\mathbb{P}(W - d \geq dt) &\leq \inf_{0 < \alpha < 1/2} \left(\frac{e^{-\alpha}}{\sqrt{1 - 2\alpha}} e^{-\alpha t} \right)^d \\
&\leq \inf_{0 < \alpha < 1/4} \exp\{d(2\alpha^2 - \alpha t)\} \quad \text{using the hint,} \\
&= e^{-dt^2/8}
\end{aligned}$$

setting $\alpha = t/4$ in the last line, which is permitted since $t < 1$. The argument to bound $\mathbb{P}(d - W \geq dt)$ is similar, and the result follows from a union bound.

(b) Let $A \in \mathbb{R}^{d \times p}$ have i.i.d. standard normal entries. Fix $u \in \mathbb{R}^p$. Use the result above to conclude that

$$\mathbb{P}\left(\left| \frac{\|Au\|_2^2}{d\|u\|_2^2} - 1 \right| \geq t \right) \leq 2 e^{-dt^2/8}.$$

   Solution: Let a_i be the ith row of A. Then $a_i^T u/\|u\|_2 \sim N(0, 1)$ and thus by independence of the rows, $\|Au\|_2^2/\|u\|_2^2$ has a χ_d^2 distribution. Now use the answer from part (a).

(c) Suppose we have (data) $u_1, \ldots, u_n \in \mathbb{R}^p$ (note each u_i is a vector), with p large and $n \geq 2$. Show that for a given $\epsilon \in (0, 1)$ and $d > 16 \log(n/\sqrt{\epsilon})/t^2$, each data point may be compressed down to $u_i \mapsto A u_i / \sqrt{d} = w_i$ whilst approximately preserving the distances between the points:

$$\mathbb{P}\left(1 - t \leq \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + t \text{ for all } i, j \in \{1, \ldots, n\}, i \neq j\right) \geq 1 - \epsilon.$$

This is the famous Johnson–Lindenstrauss Lemma.

Solution: Apply the result from (b) with $u = u_i - u_j$. Using the union bound, we get

$$\mathbb{P}\left(\left|\frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} - 1\right| \geq t\right) \leq \binom{n}{2} \times 2e^{-dt^2/8} \leq \exp\{-dt^2/8 + 2\log(n)\} \leq \epsilon.$$

In the following questions assume that X has had its columns centred and scaled to have ℓ_2-norm $\sqrt{n}$, and that Y is also centred.

9. Show that any two Lasso solutions when $\lambda > 0$ must have the same ℓ_1-norm.

   Solution: We know that for any two Lasso solutions $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$ we have that $X\hat{\beta}^{(1)} = X\hat{\beta}^{(2)}$. Thus the least squares part of their Lasso objectives will be equal. Since the two Lasso objectives must be equal (they are both Lasso solutions), the ℓ_1 parts must also be the same.

10. A *convex combination* of a set of points $S = \{v_1, \ldots, v_m\} \subseteq \mathbb{R}^{d'}$ is any point of the form

$$\alpha_1 v_1 + \cdots + \alpha_m v_m,$$

where $\alpha_j \in \mathbb{R}$ and $\alpha_j \geq 0$ for $j = 1, \ldots, m$, and $\sum_{j=1}^m \alpha_j = 1$. Carathéodory's Lemma states that if S is in a subspace of dimension d, any v that is a convex combination of points in S can be expressed as a convex combination of $d + 1$ points from S i.e. there exist $j_1, \ldots, j_{d+1} \in \{1, \ldots, m\}$ and non-negative reals $\alpha_1, \ldots, \alpha_{d+1}$ summing to 1 with

$$v = \alpha_1 v_{j_1} + \cdots + \alpha_{d+1} v_{j_{d+1}}.$$

With this knowledge, show that for any value of λ, there is always a Lasso solution with no more than n non-zero coefficients.

Solution: As the columns of X are centred, they live in an $n - 1$-dimensional subspace of $\mathbb{R}^n$. $X\hat{\beta}_\lambda^L$ is a convex combination of points in

$$\|\hat{\beta}_\lambda^L\|_1 \{\pm X_1, \ldots, \pm X_p\} \subset \mathbb{R}^n.$$

By Carathéodory's Lemma, there exist $1 \leq k_1, \ldots, k_n \leq p$ and an $\alpha \in \mathbb{R}^n$ with $\|\alpha\|_1 = 1$ such that

$$X\hat{\beta}_\lambda^L = \|\hat{\beta}_\lambda^L\|_1 \sum_{j=1}^n \alpha_j X_{k_j}.$$

Since $\|\hat{\beta}_\lambda^L\|_1 \|\alpha\|_1 = \|\hat{\beta}_\lambda^L\|_1$, the expression on the RHS must also constitute a Lasso solution with only n non-zero components.