MODERN STATISTICAL METHODS Part III
Example Sheet 2 (of 4) RDS/Michaelmas 2025

Questions 6 and 11 will be marked.

1. Consider the setup of Theorem 7 but where $(Y_1, x_1), \ldots, (Y_n, x_n)$ are i.i.d., $\mathbb{E}(\varepsilon_i \,|\, x_i) = 0$ and $\mathrm{Var}(\varepsilon_i \,|\, x_i) = \sigma^2$. Suppose each $x_i \sim U[0,1]$. Suppose further that the RKHS associated with the kernel k is such that $\|f^0\|_{\mathcal{H}} \leq \sigma$.

   (a) For the Gaussian kernel k with unit bandwidth, there exist universal constants $C_1, c_1 > 0$ such that the resulting integral operator has jth eigenvalue $\mu_j \leq C_1 \exp(-c_1 j)$ (and each has multiplicity 1). Show that for a tuning parameter choice λ_n that you should specify, there exists a universal constant $C > 0$ such that
$$\frac{1}{n}\mathbb{E}\biggl\{\sum_{i=1}^n \bigl(\hat{f}_{\lambda_n}(x_i) - f^0(x_i)\bigr)^2\biggr\} \leq C\sigma^2 \frac{\log(en)}{n}$$
   for all $n \in \mathbb{N}$.

   (b) Now consider the case where k is the second-order Sobolev kernel, where it is known that the jth eigenvalue $\mu_j \leq \{(j-1)\pi\}^{-4}$ (and each has multiplicity 1). For a tuning parameter choice λ_n that you should specify, show that there exists a universal constant $C > 0$ such that
$$\frac{1}{n}\mathbb{E}\biggl\{\sum_{i=1}^n \bigl(\hat{f}_{\lambda_n}(x_i) - f^0(x_i)\bigr)^2\biggr\} \leq C\sigma^2 n^{-4/5}$$
   for all $n \in \mathbb{N}$.

2. Consider the setup of Question 1 but where we only know $\mathrm{Var}(\varepsilon_i \,|\, x_i) \leq \sigma^2$. Show that
$$\frac{1}{n}\mathbb{E}\left(\sum_{i=1}^n (f^0(x_i) - \hat{f}_\lambda(x_i))^2 \,|\, x_1, \ldots, x_n\right) \leq \frac{\sigma^2}{\lambda}\frac{1}{n}\sum_{i=1}^n \min(d_i/4, \lambda) + \|f^0\|_{\mathcal{H}}^2 \frac{\lambda}{4n}.$$

3. In the setting of Question 2, suppose that $\sigma^2 \leq c_\sigma$ and $\|f^0\|_{\mathcal{H}}^2 \leq c_f$ for some $c_\sigma, c_f > 0$ known to the practitioner. Consider the data-driven choice of tuning parameter given by
$$\hat{\lambda} := \operatorname*{argmin}_{\lambda > 0} \left\{\frac{c_\sigma}{n}\sum_{i=1}^n \frac{d_i^2}{(d_i + \lambda)^2} + \frac{\lambda c_f}{4n}\right\}.$$
   Show that
$$\frac{1}{n}\mathbb{E}\biggl\{\sum_{i=1}^n \bigl(\hat{f}_{\hat{\lambda}}(x_i) - f^0(x_i)\bigr)^2\biggr\} \leq \frac{1}{4}\max\biggl(\frac{c_\sigma}{\sigma^2}, \frac{c_f}{\|f^0\|_{\mathcal{H}}^2}\biggr) \inf_{\gamma > 0}\biggl(\frac{\sigma^2\phi(\gamma)}{n\gamma} + \gamma\|f^0\|_{\mathcal{H}}^2\biggr),$$
   where
$$\phi(\gamma) := \sum_{j \in J} \min(4\gamma, \mu_j).$$

4. 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 $\mathbb{E}(\psi^\top x \psi^\top x') = x^\top x'$. Construct a random feature map $\hat{\phi} : \mathbb{R}^p \to \mathbb{R}$ such that $\mathbb{E}\{\hat{\phi}(x)\hat{\phi}(x')\} = (x^\top x')^2$.

5. 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).

6. (a) Let $W \sim \chi^2_d$. Show that
$$\mathbb{P}(|W/d - 1| \geq t) \leq 2e^{-dt^2/8}$$
for $t \in (0, 1)$. You may use the facts that the mgf of a χ^2_1 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 $|\alpha| < 1/4$.

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

   (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 $u_i \mapsto Au_i/\sqrt{d} = w_i$ 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 result is known as the *Johnson–Lindenstrauss Lemma.*

In the following questions, assume that $X \in \mathbb{R}^{n \times p}$ is a column-centred matrix of predictors with each column additionally scaled to have the same ℓ_2-norm.

7. 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 given a column-centred matrix of predictors $X \in \mathbb{R}^{n \times p}$ and response $Y \in \mathbb{R}^n$, for any value of $\lambda \geq 0$, there is always a Lasso solution with no more than n non-zero coefficients.

8. Show that when the columns of X are orthogonal (so necessarily $p \leq n$) and scaled to have ℓ_2-norm $\sqrt{n}$, the kth component of the Lasso estimator is given by

$$\hat{\beta}^{\mathrm{L}}_{\lambda,k} = (|\hat{\beta}^{\mathrm{OLS}}_k| - \lambda)_+ \mathrm{sgn}(\hat{\beta}^{\mathrm{OLS}}_k)$$

where $(\cdot)_+ = \max(0, \cdot)$. What is the corresponding estimator if the ℓ_1 penalty $\|\beta\|_1$ in the Lasso objective is replaced by the ℓ_0 penalty $\|\beta\|_0 := |\{k : \beta_k \neq 0\}|$?

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

10. Show that if $\lambda \geq \lambda_{\max} := \|X^\top Y\|_\infty / n$, then $\hat{\beta}^{\mathrm{L}}_\lambda = 0$.

11. When proving the theorems on the prediction error of the Lasso, we started with the so-called basic inequality that

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Show that in fact we can improve this to

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leq \frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

12. Consider a setup where the Lasso solution is unique for all $\lambda > 0$. Let $\hat{\beta}^{\mathrm{L}}_{\lambda_1}$ and $\hat{\beta}^{\mathrm{L}}_{\lambda_2}$ be two Lasso solutions at different values of the regularisation parameter. Suppose that $\mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_{\lambda_1}) = \mathrm{sgn}(\hat{\beta}^{\mathrm{L}}_{\lambda_2})$. Show that then for all $t \in [0, 1]$,

$$t\hat{\beta}^{\mathrm{L}}_{\lambda_1} + (1-t)\hat{\beta}^{\mathrm{L}}_{\lambda_2} = \hat{\beta}^{\mathrm{L}}_{t\lambda_1 + (1-t)\lambda_2}.$$

[*Hint: Check the KKT conditions.*] Conclude that the solution path $\lambda \mapsto \hat{\beta}^{\mathrm{L}}_\lambda$ is piecewise linear with a finite number of knots (points λ where the solution path is not linear at λ) and these occur when the sign of the Lasso solution changes.