Problem 1

Prove the following results. These were stated in class without a formal proof.

(a) Let $(x_i)_{i=1}^n \subset [0,1]^d$ be a set of deterministic locations and let $(y_i)_{i=1}^n$ be n i.i.d random variables such that

$$y_i = f(x_i) + \epsilon_i, \text{ with } \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

Let $\hat{f} : [0,1]^d \to \mathbb{R}$ be a nonparametric estimator of the regression function f and show that for a fixed $x \in [0,1]^d$, the Mean Squared Error decouples as:

$$\mathrm{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] = \left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)^2 + \mathrm{var}\left(\hat{f}(x)\right)$$

Note. The terms $\left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)^2$ and $\mathrm{var}\left(\hat{f}(x)\right)$ are respectively the squared bias and variance of the estimator $\hat{f}(x)$.

(b) Suppose now we are given a new independent sample $(x_{\text{new}}, y_{\text{new}})$. Show how the prediction error

$$\mathrm{E}\left[\left(y_{\text{new}} - \hat{f}(x_{\text{new}})\right)^2\right]$$

is related to the bias and variance terms of the estimator $\hat{f}$ computed from $(x_i, y_i)_{i=1}^n$.

(c) For a non-negative random variable L, verify that if $\mathrm{E}[L] \le a$ then

$$P\left(\frac{L}{a} > \frac{1}{\epsilon}\right) \le \epsilon$$

If you use Markov's inequality to show this, please reprove Markov's inequality (it's short...)

(d) Consider now a random design, i.e. $x_i \overset{iid}{\sim} F$, with x_i taking values in $\mathbb{R}^p$. Let F be such that $\mathrm{E}[x_i] = 0$, and $\mathrm{var}(x_i) = \Sigma$. Assume the linear model

$$y_i = x_i^\top \beta + \epsilon_i$$

with $\epsilon_i \overset{iid}{\sim} G$ with $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{var}(\epsilon_i) = \sigma^2$. And assume the xs and the ϵs are independent. Show that

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to \mathrm{N}\left(0, \sigma^2 \Sigma^{-1}\right)$$

where $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y$ (hint use Slutsky's theorem).

Discuss what this formally implies about the rate of convergence of $\hat{\beta}$ to β (we talked about this in class).

(Optional) What would have been different if instead we had a fixed design, i.e. $\{x_i\} \subset \mathbb{R}^p$ were deterministic quantities.

Problem 2

This problem is about conducting a basic simulation study (in `R`) comparing parametric and non-parametric rates. Suppose $x_i \overset{iid}{\sim} U[-1,1]$, and

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\epsilon_i \sim N(0,1)$. For different f, we will explore the appropriateness of parametric vs non-parametric methods.

For each of the following, compute the empirical MSE $\frac{1}{n}\sum_i \left(\hat{f}(x_i) - f(x_i)\right)^2$, where $\hat{f}$ is estimated by (i) linear regression; (ii) parametric polynomial regression on polynomials (in x) of degrees 2 to 5; (iii) Nadaraya-Watson estimation with a "box" kernel (see `help("ksmooth")`), and (iv) Nadaraya-Watson with a "gaussian" kernel (see `help("ksmooth")`). For both NW estimators let the bandwidth be $h = n^{-\frac{1}{5}}$.

(a) $f(x) = 2x$.

(b) $f(x) = sin(x * \pi)$.

(c) $f(x) = 2x + x^3 - 6x^4$.

(d) $f(x) = \frac{1}{1+(5x)^2}$.

For each of these, calculate the MSE for varying values of n for each estimator. Make appropriate plot(s) to compare these estimators. Give a short writeup stating comparisons/conclusions, with particular focus on choices (c) and (d).

The R commands `replicate`, `poly`, `lm`, `predict`, `rnorm`, and `runif`, `ksmooth` might come in handy.