Choose 3 of the 4 problems on this problem set to do (though, you should also feel free to choose all 4 if you want more practice!)

Problem 1

Prove the following results. These were stated in class (some were proved in class — try to prove them again without referring to the video/notes!).

(a) Given an estimator $\hat{\theta}$ of a fixed parameter θ show that the MSE of $\hat{\theta}$ decouples as:

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

Note. This is the bias-variance tradeoff.

(b) For a non-negative random variable T, verify that if $\mathrm{E}\left[T\right] \leq a$ then

$$P\left(\frac{T}{a} > \frac{1}{\epsilon}\right) \leq \epsilon$$

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

(c) If $x_i \overset{iid}{\sim} F$ (with each $x_i \in \mathbb{R}^p$) with $\mathrm{E}\left[x_i\right] = 0$, and $\mathrm{var}\left(x_i\right) = \Sigma$; and

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

with $\epsilon_i \overset{iid}{\sim} G$ with $\mathrm{E}\left[\epsilon_i\right] = 0$, $\mathrm{var}\left(\epsilon_i\right) = \sigma^2$. 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).

Problem 2

This problem is about conducting a basic simulation study (in `R`) comparing parametric and non-parametric rates. Suppose $x_i \overset{iid}{\sim} U[0,1]$, and

$$y_i = f\left(x_i\right) + \epsilon_i$$

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, compare rates for $\frac{1}{n}\sum_i \left(\hat{f}(x_i) - f(x_i)\right)^2$ 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, and (iv) Nadaraya-Watson with a "gaussian" kernel. For both NW estimators make an appropriate choice of bandwidth (and defend this choice).

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

Problem 2 continued on next page...

(b) $f(x) = sin(2\pi x)$.

(b) $f(x) = sin(30x)$.

For each of these, calculate MSE for varying values of n for each estimator. Make appropriate plot(s) to compare these estimators. Give a short writeup stating comparisons/conclusions.

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

Problem 3

This problem is related to non-parametric estimation using the box-kernel (which we saw in class). Consider the triangular array of pairs $\{(x_{i,n},\, y_{i,n})\}_{i\leq n,\, n=1,\ldots}$. Suppose $x_{i,n} = i/n$ (so, $x_{i,n}$ is *deterministic*), and $y_{i,n}$ is generated as

$$y_{i,n} = f\left(x_{i,n}\right) + \epsilon_{i,n}$$

with unknown f, where $\epsilon_{i,n}$ are distributed iid with $\mathrm{E}\left[\epsilon_{i,n}\right] = 0$, $\sigma_\epsilon^2 \equiv \mathrm{var}\left(\epsilon_{i,n}\right) < \infty$. In this problem we consider the task of estimating f.

(a) Suppose we would like to estimate f by a piecewise constant function. Fix n, and suppose we observe a single column of our triangular array $\left(x_{1,n}, y_{1,n}\right), \ldots, \left(x_{n,n}, y_{n,n}\right)$. For a fixed integer $0 < k(n) < n$, consider the class of piecewise constant functions on $(0, 1]$ with $k(n)$ equally spaced break-points/knots:

$$\mathcal{F}_n = \left\{ f : (0,1] \to \mathbb{R},\text{ with } f\left(x\right) \equiv \sum_{i=1}^{k(n)} c_i I\left\{\frac{i-1}{k(n)} < x \leq \frac{i}{k(n)}\right\} \;\middle|\; c_1, \ldots, c_{k(n)} \in \mathbb{R} \right\}$$

If $\epsilon_{i,n}$ were drawn iid from $N\left(0, \sigma_\epsilon^2\right)$, show that the element of $\mathcal{F}_n$ that maximizes the likelihood is given by

$$\hat{f}_n(x) = \sum_{j=1}^{k(n)} \left(\frac{\sum_{i=1}^{n} y_{i,n} I\left\{x_{i,n} \in A_{j,n}\right\}}{\sum_{i=1}^{n} I\left\{x_{i,n} \in A_{j,n}\right\}} \right) I\left\{x \in A_{j,n}\right\}$$

for each $x \in (0, 1]$ with $A_{j,n} = \left(\frac{j-1}{k(n)}, \frac{j}{k(n)}\right]$. (Hint: draw a picture! The notation is dense, but the problem is relatively straightforward)

(b) As we observe more data, we likely want to increase the number of knots/breakpoints in our approximation. Consider n increasing, and $k(n)$ increasing (in n) with $k(n) \leq n$ (for each n). Suppose that f is differentiable, and that there exists $c \in \mathbb{R}$, such that for all x, $|f'(x)| \leq c$. For a fixed $x_0 \in [0, 1)$ verify that the mean square error (MSE) of $\hat{f}_n\left(x_0\right)$ decreases like

$$\mathrm{E}\left[\left(\hat{f}_n\left(x_0\right) - f\left(x_0\right)\right)^2\right] = O\left(\frac{\sigma_\epsilon^2 k(n)}{n} + \frac{c^2}{k(n)^2}\right)$$

(Hint: look at the bias and variance).

(c) What does (b) tell you about the rate of convergence of your estimator (with an optimal choice of $k(n)$)? How does this bound compare to the rate of convergence you would get using a box kernel?

Problem 4

For this problem, we use the `fev` data discussed in class. Let y be `fev` and the covariate be `height`. Fit a kernel regression estimate of the form $y = f(x) + \epsilon$.

(a) For a rectangular kernel, use leave-one-out cross validation and five-fold cross validation to choose the bandwidth. Compare the choices of bandwidth and the resulting fits.

(b) Repeat **(a)** using a Epanechnikov kernel.