Problem Set 2 - Pedro Henrique Cavalcanti Rocha

Question 1

We know that the function **a** is of the order $\mathcal{O}(h^k)$ if and only if:

$$\frac{a(h)}{h^k} \longrightarrow t, t \in \mathbb{R}, h \longrightarrow 0$$

Thus, take $h^* = (\frac{1}{n})^{0.2}$
Now we have:

$$\sqrt{nh^*} = \sqrt{n(\frac{1}{n})^{0.2}} = \sqrt{\frac{n}{n^{0.2}}} = \sqrt{n^{0.8}} = (n^{0.8})^{0.5} = n^{0.4}$$

Hence, $\sqrt{nh^*}$ is of order $\mathcal{O}(n^{0.4})$

Question 2

We know that the function a is said to be $o(h^k)$ if and only if:

$$\frac{a(h)}{h^k} \longrightarrow t, t \in \mathbb{R}, h \longrightarrow 0$$

i.e, $a(h)$ goes to zero "way faster" than h^k.
Then given the functions f and g of order $o(n^2)$, $o(n)$, respectively, we have:

$$\frac{f(n)}{n^2} \longrightarrow t, t \in \mathbb{R}$$

$$\frac{g(n)}{n} \longrightarrow t, t \in \mathbb{R}$$

Thus, we have:

$$\frac{f(n)}{n^2}\frac{g(n)}{n} = \frac{f(n)g(n)}{n^3} \longrightarrow t, t \in \mathbb{R}$$

Hence, fg is of order $o(n^3)$.

Question 3

First, we define the Integrated Squared Error (ISE):

$$ISE(h) = \int (\hat{f}(x_0) - f(x_0))^2 dx_0$$

The Mean Integrated Squared Error (MISE) is:

$$MISE(h) = E\left(\int (\hat{f}(x_0) - f(x_0))^2 dx_0\right) = \int MSE(\hat{f}(x_0)) dx_0$$

Rewriting the MISE, we have:

$$MISE(h) = \int MSE(\hat{f}(x_0))dx_0 = \int b(x_0)^2 + Var(\hat{f}(x_0))dx_0$$

The bias term is:

$$b(x_0) = \frac{1}{2}h^2 f''(x_0) \int z^2 K(z)dz + \mathcal{O}^3$$

And the variance is:

$$Var(\hat{f}(x_0)) = \frac{1}{nh} f(x_0) \int K(z)^2 dz + o(\frac{1}{nh})$$

By disregarding the asymptotic terms, we can rewrite the MISE as:

$$MISE(h) = \int (\frac{1}{2} f''(x_0) \int z^2 K(z)dz)^2 + \frac{1}{nh} f(x_0) \int K(z)^2 dz$$

$$MISE(h) = \int \frac{1}{4} h^4 f''(x_0)^2 \left(\int z^2 K(z)dz \right)^2 dx_0 + \int \frac{1}{nh} f(x_0) \int K(z)^2 dz dx_0$$

$$MISE(h) = \int \frac{1}{4} h^4 f''(x_0)^2 \left(\int z^2 K(z)dz \right)^2 dx_0 + \frac{1}{nh} \int f(x_0)dx_0 \int K(z)^2 dz$$

Since $\int f(x_0)dx_0 = 1$, then:

$$MISE(h) = \int \frac{1}{4} h^4 f''(x_0)^2 \left(\int z^2 K(z)dz \right)^2 dx_0 + \frac{1}{nh} \int K(z)^2 dz$$

$$MISE(h) = \frac{1}{4} h^4 \left(\int f''(x_0)^2 dx_0 \right) \left(\int z^2 K(z)dz \right)^2 + \frac{1}{nh} \int K(z)^2 dz$$

The globally optimal bandwidth h^* minimizes MISE(h). Then the FOC is:

$$h^3 \left(\int f''(x_0)^2 dx_0 \right) \left(\int z^2 K(z)dz \right)^2 - \frac{1}{nh^2} \int K(z)^2 dz = 0$$

$$h^5 \left(\int f''(x_0)^2 dx_0 \right) \left(\int z^2 K(z)dz \right)^2 = \left(\frac{1}{n} \int K(z)^2 dz \right)$$

$$h^5 = n^{-1} \left(\int f''(x_0)^2 dx_0 \right)^{-1} \left(\int z^2 K(z)dz \right)^{-2} \left(\int K(z)^2 dz \right)$$

$$h^* = n^{-0.2} \delta \left(\int f''(x_0)^2 dx_0 \right)^{-0.2}$$

where δ is as it is defined in Cameron et Trived (2005).

Question 4

First, we can rewrite the cross-validation function:

$$CV = \frac{1}{n}\sum(y_i - \bar{m}_{-i}(x_i,h))^2$$

$$= \frac{1}{n}\sum(y_i - \bar{m}_{-i}(x_i,h) + m(x_i) - m(x_i))^2$$

$$= \frac{1}{n}\sum(\varepsilon_i - \bar{m}_{-i}(x_i,h) + m(x_i))^2$$

$$= \frac{1}{n}\sum(\varepsilon_i^2 + 2(m(x_i) - \bar{m}_{-i}(x_i,h))\varepsilon_i + (m(x_i) - \bar{m}_{-i}(x_i,h))^2)$$

Taking the expectation is:

$$\mathbb{E}_{\mathbf{x},\varepsilon}(CV) = \mathbb{E}_{\mathbf{x},\varepsilon}[\frac{1}{n}\sum(\varepsilon_i^2 + 2(m(x_i) - \bar{m}_{-i}(x_i,h))\varepsilon_i + (m(x_i) - \bar{m}_{-i}(x_i,h))^2)]$$

Since we are assuming the data is i.i.d, we have:

$$= \frac{1}{n}\sum \mathbb{E}_{\mathbf{x},\varepsilon}[(\varepsilon_i^2 + 2(m(x_i) - \bar{m}_{-i}(x_i,h))\varepsilon_i + (m(x_i) - \bar{m}_{-i}(x_i,h))^2)]$$

Redistributing:

$$= \mathbb{E}_{\mathbf{x},\varepsilon}[\varepsilon_i^2] + 2\mathbb{E}_{\mathbf{x},\varepsilon}[(m(x_i) - \bar{m}_{-i}(x_i,h))\varepsilon_i] + \mathbb{E}_{\mathbf{x},\varepsilon}[(m(x_i) - \bar{m}_{-i}(x_i,h))^2].$$

From previous discussions (statistics course), given the data is i.i.d, we have the following property:

$$2\mathbb{E}_{\mathbf{x},\varepsilon}[(m(x_i) - \bar{m}_{-i}(x_i,h))\varepsilon_i] = 0$$

Hence:

$$\mathbb{E}_{\mathbf{x},\varepsilon}[CV] = \mathbb{E}_{\mathbf{x},\varepsilon}[\varepsilon_i^2] + \mathbb{E}_{\mathbf{x},\varepsilon}[(m(x_i) - \bar{m}_{-i}(x_i,h))^2]$$

$$= \varepsilon^2 + \mathbb{E}_{\mathbf{x},\varepsilon}[((\bar{m}_{-i}(x_i,h) - m(x_i))^2]$$

This resembles a function of type $f(x) = a + bg(x)$. Thus we can minimize only in terms of the function $g(x)$.

Now we have to prove that minimizing our $g(.)$ is equivalent to minizing the Integrated Mean Squared Error (IMSE).

(Not really sure how to go from here ...)