Kernel density estimation

17 marks (undergrads) (plus 4 bonus available)

21 marks (grads)

Kernel density estimation.

a. Recall from the notes the general results on the bias (up to $O(h^4)$) and variance (up to order $O(\frac{h}{n})$) of a "simplified kernel" density estimator.

   i. **(4 marks)** Prove that the following kernel is a "simplified kernel" (in the sense of the notes).

$$K(w) = \begin{cases} \frac{15}{32}(1-w^2)(3-7w^2) & w \in [-1,1] \\ \\ 0 & \text{elsewhere.} \end{cases}$$

we'll show $\int K(w)dw = 1 \quad \int wK(w)dw = 0 \quad \textit{and } 0 \leq \int w^2 K(w)dw < \infty$

$$\begin{aligned}\int K(w)dw &= \int_{-\infty}^{\infty} \frac{15}{32}(1-w^2)(3-7w^2)dw \\ &= \frac{15}{32}\int_{-1}^{1}(7w^4 - 10w^2 + 3)dw \\ &= \frac{15}{32}(\frac{7}{5}w^5 - \frac{10}{3}w^3 + 3w)|_{w=-1}^{1} \\ &= 1\end{aligned}$$

$$\begin{aligned}\int wK(w)dw &= \int_{-\infty}^{\infty} \frac{15}{32}(1-w^2)(3-7w^2)wdw \\ &= \frac{15}{32}\int_{-1}^{1}(7w^5 - 10w^3 + 3w)dw \\ &= \frac{15}{32}(\frac{7}{6}w^6 - \frac{10}{4}w^4 + \frac{3}{2}w^2)|_{w=-1}^{1} \\ &= 0\end{aligned}$$

$$\begin{aligned}\int w^2K(w)dw &= \int_{-\infty}^{\infty} \frac{15}{32}(1-w^2)(3-7w^2)w^2dw \\ &= \frac{15}{32}\int_{-1}^{1}(7w^6 - 10w^4 + 3w^2)dw \\ &= \frac{15}{32}(\frac{7}{7}w^7 - \frac{10}{5}w^5 + \frac{3}{3}w^3)|_{w=-1}^{1} \\ &= 0\end{aligned}$$

and we are done

ii. **(8 marks)** Determine the (approximate) mean squared error (from the slides) of $\tilde{f}_K(x)$ for the above kernel K and for arbitrary $f(x)$.

mse=bias+variance

$$\begin{aligned} bias[\tilde{f}_K(x)] &= \frac{1}{2}\sigma_k^2 h^2 f^{''}(x) + \frac{1}{6}h^3 f^{'''}(x)\int w^3 K(w)dw + O(h^4) \\ &= 0 + \frac{1}{6}h^3 f^{'''}(x)(\frac{7}{8}w^8 - \frac{10}{6}w^6 + \frac{3}{4}w^4)|_{w=-1}^{1} + O(h^4) \\ &= 0 + O(h^4) \end{aligned}$$

for variance we need to evaluate $\int K^2(w)dw$

$$\begin{aligned} \int K^2(w)dw &= \int_{-\infty}^{\infty} (\frac{15}{32})^2(1-w^2)^2(3-7w^2)^2 dw \\ &= (\frac{15}{32})^2 \int_{-1}^{1}((1-2w^2+w^4)(49w^4-42w^2+9))dw \\ &= \frac{5}{4} \end{aligned}$$

$$\begin{aligned} var[\tilde{f}_K(x)] &= \frac{1}{nh}f(x)\int K^2(w)dw - \frac{1}{n}[f(x)]^2 + O(\frac{h^2}{n}) \\ &= \frac{1}{nh}f(x) * \frac{5}{4} - \frac{1}{n}[f(x)]^2 + O(\frac{h^2}{n}) \\ &= \frac{1}{nh}f(x) * \frac{5}{4} - \frac{1}{n}[f(x)]^2 + O(\frac{h}{n}) \text{ as h<1, the question ask for } O(\frac{h}{n}) \end{aligned}$$

$$\begin{aligned} mse &= \frac{1}{nh}f(x) * \frac{5}{4} - \frac{1}{n}[f(x)]^2 + O(\frac{h}{n}) + (O(h^4))^2 \\ &= \frac{1}{n}(\frac{5}{4h}f(x) - f^2(x)) + O(\frac{h}{n}) + O(h^8) \end{aligned}$$

iii. **(5 marks)** Determine the case for the above K when the true underlying density $f(x)$ is $N(0,1)$.
we just put $f_{N(0,1)}$ in the mse equation,

$$mse = \frac{5}{4hn}(\frac{1}{\sqrt{2\pi}}e^{-0.5x^2}) - \frac{1}{n}(\frac{1}{2\pi}e^{-x^2}) + O(\frac{h}{n}) + O(h^8)$$

b. **Graduate students (bonus undergraduates) (4 marks)**

Consider the general ASH estimate (non-naive ASH):

$$\widehat{f}(x,m) \quad = \quad \frac{1}{nh}\sum_{|i|<m} w_m(i)v_{k+i}$$

for $x \in B_k$. Here the weights $w_m(i) \geq 0$ and the intervals $B_j = [b_j, b_{j+1})$ indexed by $j = 0, \pm1, \pm2, \pm3, \ldots$ partition the entire real line.

The intervals are each of width $(b_{j+1} - b_j) = \frac{h}{m}$ and v_j is the number of xs in B_j. The total sample size is $n = \sum_{|j|=0}^{\infty} v_j$.

Prove that if $\sum_{|i|<m} w_m(i) = m$ then $\int \widehat{f}(x,m)dx = 1$.