Homework – 2

Exercise 1:[Simulation experiments and KDE] Here is list of steps you have to perform with Python. At each step, you are expected to provide comments/explanations about the phenomenon of interest as well as graphs which illustrate your conclusions. Do not forget indicating legends and titles on the graph axes.

1. Draw one realization of $n = 100$ *independent* random variables $X_1, \ldots, X_n$ from a standard Gaussian distribution $\mathcal{N}(0, 1)$.
   (a) For a kernel $K = \mathbb{1}_{[-1,1]}/2$, compute and display the graph of the resulting KDE obtained with different bandwidth values $h \in \mathcal{H} = \{10^{-5}, 10^{-3}, 10^{-1}, 10\}$. What do you see?
   (b) Repeat the same experiment with $n = 10\,000$. Is there a change? Comment.
   (c) Let us now consider a Gaussian kernel $K'(x) = 1/\sqrt{2\pi} e^{-x^2/2}$, $x \in \mathbb{R}$. Reproduce the same experiments as in the above two questions and make a "by eye" comparison regarding the quality of the approximation you get.
2. Using a Monte-Carlo (MC) strategy, compute (an approximation to) the $MSE_h(x_0)$ criterion for the KDE built from K, $n = 100$, and varying the bandwidth on the grid $\mathcal{H}$.
   (a) Start with , with $x_0 = -2$ Which value of h is the best?
   (b) Repeat the experiments with $x_0 = 0.1$. Same question. Is there a difference? What conclusion could you draw?
   (c) Let us now consider a somewhat different criterion denoted by $DMSE(h)$ defined by
   $$DMSE(h) = \frac{1}{T} \sum_{t=1}^{T} MSE_h(x_t),$$
   with $x_t = -3 + t \times 6/N$, for $1 \leq t \leq T$ and $T = 200$.
   Display the graph of $h \in \mathcal{H} \mapsto DMSE(h)$. Which value of h is the best? Is there a change compared to what you observed for the $MSE_h(x_0)$?
3. Draw now a realization of $n = 1\,000$ independent random variables such that the first 200 ones and from a standard Gaussian and the remaining 800 ones are drawn from a $\mathcal{N}(4, \sigma^2)$, with $\sigma^2 = 8$.
   (a) Draw the graphs of the KDE function obtained with K for $h \in \mathcal{H}$.
   (b) Draw the graph of $h \in \mathcal{H} \mapsto DMSE(h)$, with $x_t = -2 + t \times 10/N$, for $1 \leq t \leq T$ and $T = 1000$. Say which bandwidth value is the best one? Could you provide a tentative explanation?

□

Exercise 2:[Bernstein's condition and KDE] The purpose of the present exercise is to prove a concentration bound with high probability on the kernel density estimator (KDE) evaluated at x_0.

Bernstein's condition $BC(v, c)$: We say that a real-valued random variable X satisfies the Bernstein condition ($BC(v, c)$) for two constants $v, c > 0$ if, for all integers $k \geq 2$,

$$\mathbb{E}\left[|X|^k\right] \leq \frac{k!}{2} v \cdot c^{k-2}.$$

Fact: For any sample of n real-valued random variables $X_1, \ldots, X_n$ such that each X_i satisfies $BC(v, c)$, it holds for every $y > 0$ that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i > y\right] \vee \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i < -y\right] \leq e^{-\frac{ny^2}{2(v+cy)}}. \qquad \text{(BCIneq)}$$

Let $D = \{X_1, \ldots, X_n\}$ be n *independent* real-valued random variables from a probability distribution P with density f with respect to the Lebesgue measure on $\mathbb{R}$. We also recall that for a given non-negative kernel K (symmetric) and a bandwidth $h > 0$, the KDE is defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right), \qquad \forall x \in \mathbb{R}.$$

Assume that the kernel K is bounded that is, $\|K\|_\infty < +\infty$ and consider $x_0 \in \mathbb{R}$ fixed.

1. Setting $\zeta_1 = K_h(X_1 - x_0) - \mathbb{E}\left[K_h(X_1 - x_0)\right]$, prove that, for any integer $k \geq 2$, we have

$$\mathbb{E}\left[|\zeta_1|^k\right] \leq \frac{k!}{2} v \cdot c^{k-2}.$$

   where $c = 2\|K\|_\infty / h$ and $v = \mathrm{Var}(K_h(X_1 - x_0))$.

2. Deduce that, for every $t > 0$,

$$\mathbb{P}\left[\left|\hat{f}_h(x_0) - \mathbb{E}_D\left[\hat{f}_h(x_0)\right]\right| > t\right] \leq 2e^{-\frac{nt^2}{2(v+ct)}}.$$

3. Prove that the next two statements are equivalent

   (a) For every $y > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i > y\right] \leq e^{-\frac{ny^2}{2(v+cy)}},$$

   (b) For every $x > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i > \sqrt{\frac{2vx}{n}} + \frac{cx}{n}\right] \leq e^{-x}.$$

4. Deduce that, for every $x > 0$,

$$\mathbb{P}\left[\left(\hat{f}_h(x_0) - \mathbb{E}_D\left[\hat{f}_h(x_0)\right]\right)^2 > 2\frac{2vx}{n} + 2\left(\frac{cx}{n}\right)^2\right] \leq e^{-x},$$

   where you can use that $(a+b)^2 \leq 2(a^2 + b^2)$, for all $a, b \in \mathbb{R}$.

5. Recalling that the kernel is bounded, justify why you could apply instead Hoeffding's inequality to $\hat{f}_h$.

6. After applying Hoeffding's inequality, compare the resulting upper bound to the former one derived from $BC(v, c)$. Which one is the tightest? Why?

$\square$