STAT 824 hw 01

Hoeffding's inequality, KS test, Brownian bridge, kernel density estimation, Hölder smoothness, higher order kernels, CV for KDE bandwidth selection

1. Let B be a Brownian motion and B_0 be a Brownian bridge. Show $\operatorname{Var} \int_0^1 B(t)dt = 1/3$.

2. (a) Use Hoeffding's inequality to show that for $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$ we have $P(|\bar{X}_n - p| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$ for every $\varepsilon > 0$.

   (b) Let $Y_1, \dots, Y_n$ be iid with continuous cdf F_Y. Show that $P(|\hat{F}_n(y) - F(y)| \leq \varepsilon) \geq 1 - 2e^{-2n\varepsilon^2}$ for each $y \in \mathbb{R}$, where $\hat{F}_n(y) = n^{-1} \sum_{i=1}^n \mathbf{1}(Y_i \leq y)$.

   (c) Explain how what you proved in part (b) is different from the DKW inequality.

3. For any random variable $Y \in [a, b]$, show that $\operatorname{Var} Y \leq (b - a)^2/4$.

4. Consider the random variables X and Y, where

$$Y \sim \text{Normal}(0, 1)$$
$$X|\delta \sim \text{Normal}(a \cdot \delta, 1 - a^2), \text{ where } \delta \in \{-1, 1\}, \text{ with } P(\delta = 1) = 1/2,$$

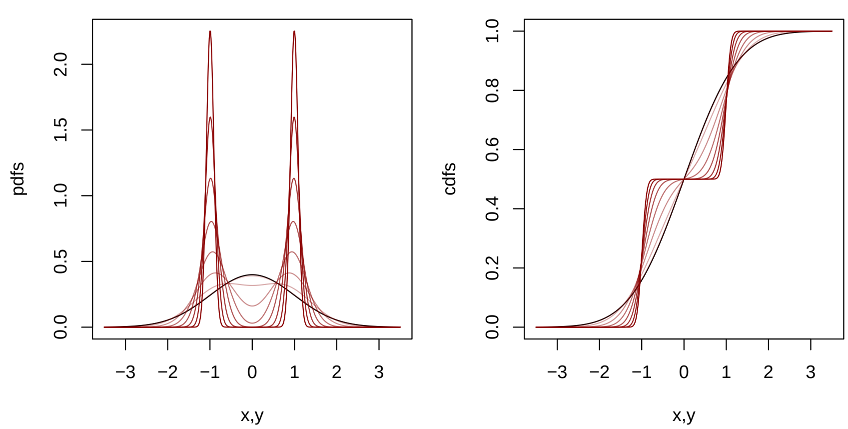
   for some $a > 0$. Suppose random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ are drawn.

   (a) Give $\mathbb{E}X$.

   (b) Give $\operatorname{Var} X$.

   (c) Give $\mathbb{E}Y^3$ and $\mathbb{E}X^3$.

   (d) Fix $n = 60$ and $m = 80$ and, for each $a \in \{1 - (1/2)^j : j = 0, 1, \dots, 8\}$, generate 100 random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ (the densities are pictured below) and report for each value of a the proportion of times the two-sample Kolmogorov-Smirnov test rejects the null hypotheses of equal cdfs (make a table). Do NOT use the R function `ks.test()`; write your own code in R or python and turn it in along with the table.



   (e) Generate a large number of Brownian bridges in order to get approximations to the 0.70, 0.80, 0.90, 0.95, and 0.99 quantiles of the distribution with cdf $\text{KS}(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i+1}e^{-2i^2x^2}$. Turn in a table of these values.

5. Given a random sample $X_1, \ldots, X_n$, find $\int_{\mathbb{R}} x \hat{f}_n(x) dx$ and $\int_{\mathbb{R}} x^2 \hat{f}_n(x) dx$ when

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $h > 0$ and

(a) K the standard Normal density.

(b) K is a kernel of order 2.

(c) What is the effect of these different kernels on $\int_{\mathbb{R}} x^2 \hat{f}_n(x) dx - (\int_{\mathbb{R}} x \hat{f}_n(x) dx)^2$?

6. Consider the function

$$N(x) = \begin{cases} \frac{32}{3} u^3, & 0 \le u < 1/4 \\ 32u^2 - 32u^3 - 8u + 2/3, & 1/4 \le u < 2/4 \\ 32(1-u)^2 - 32(1-u)^3 - 8(1-u) + 2/3, & 2/4 \le u < 3/4 \\ \frac{32}{3}(1-u)^3, & 3/4 \le u < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Identify β and L such that $N \in \mathcal{H}(\beta, L)$.

7. Let $\{\varphi_m(\cdot)\}_{m=0}^{\infty}$ represent polynomials defined by

$$\varphi_0(u) = \frac{1}{\sqrt{2}}, \quad \varphi_m(u) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{d^m}{du^m}\left[(u^2 - 1)^m\right], \quad m = 1, 2, \ldots$$

for $u \in [-1, 1]$. These are known as the Legendre polynomials on $[-1, 1]$; they are *orthonormal with respect to the Lebesgue measure*, which means they have the property

$$\int_{-1}^{1} \varphi_m(u) \varphi_k(u) du = \begin{cases} 1, & m = k \\ 0, & m \ne k. \end{cases}$$

Proposition 1.3 of [2] gives that the function $K : \mathbb{R} \to \mathbb{R}$ given by

$$K(u) = \sum_{m=0}^{\ell} \varphi_m(0) \varphi_m(u) \mathbf{1}(|u| \le 1) \tag{1}$$

is a kernel of order ℓ.

(a) Use (1) to construct a kernel of order 1.

(b) Use (1) to construct a kernel of order 2.

(c) Generate some data and make a plot of the KDE (you must code your own KDE—no using built-in functions) based on these two kernels. Include in the plot the true density from which the data were generated (it is up to you how you generate the data. Be creative!). Turn in your code along with the plots.

8. Write an (elegant) R or python function that chooses the bandwidth $h = \text{argmin}_{h>0} CV(h)$ from a grid of candidate values. Refer to Lec 2. Make your own choice of the kernel function $K(u)$. Your function should have two arguments: x for the data values and N for the number of candidate bandwidths in the grid.
   (a) Include your R or python function when you turn in your hw.
   (b) Generate some data from a distribution of your choice and plot on a single set of axes
      1. the true density,
      2. the KDE under the leave-one-out CV bandwidth,
      3. and the KDE under the Sheather-Jones bandwidth.

References

[1] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[2] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.