Homework Assignment 3

SDS 384-11 Theoretical Statistics
Deadline: March 26th
Please do not add your name to the HW submission.
Also do not add collaborators here or in the comments section of Canvas.

1. In this question we consider the Jackknife estimate of variance of a symmetrical measurable function of $n-1$ variables S. Let $X_1, \ldots, X_n - 1$ be i.i.d. Consider $S = S(X_1, \ldots, X_{n-1})$. Now let

$$S_i = S(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$$

So $S = S_n$. If S has finite variance, then the Jackknife estimate of its variance is given by:

$$\text{var}_{JACK}(S) = \sum_i \left(S_i - \frac{\sum_j S_j}{n} \right)^2$$

In Efron and Stein's Annals of Statistics paper in 1981 the following remarkable result was proven.

$$\text{var}(S) \le E\left(\text{var}_{JACK}(S)\right) \tag{1}$$

This is what we will prove here today. First define $V_i = E[S|X_1, \ldots, X_i] - E[S|X_1, \ldots, X_{i-1}]$.

   (a) Prove that $\text{var}(S) = \sum_{i=1}^{n-1} EV_i^2$

   (b) Prove that $E\text{var}_{JACK}(S) = (n-1)E[(S_1 - S_2)^2]/2$

   (c) Now prove Eq 1.

2. In this question we will look at the Gaussian Lipschitz theorem. Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(0,1)$.

   (a) Prove that the order statistics are 1-Lipschitz.

   (b) Now show that, for large enough n,

$$c\sqrt{\log n} \le E[\max_i X_i] \le \sqrt{2 \log n}$$

   where c is some universal constant.

      i. For the upper bound, let $Y = \max_i X_i$. First show that $\exp(tE[Y]) \le \sum_i E\exp(tX_i)$. Now pick a t to get the right form.

      ii. For the lower bound, do the following steps.

         A. Show that $E[Y] \ge \delta P(Y \ge \delta) + E[\min(Y, 0)]$

         B. Now show that $E[\min(Y, 0)] \ge E[\min(X_1, 0)]$

C. Finally, relate $P(Y \geq \delta)$ to $P(X_1 \geq \delta)$ by using independence.

D. Now show that $P(X_1 \geq \delta) \geq \exp(-\delta^2/\sigma^2)/c$, for some universal constant c.

E. Choose the parameter δ carefully to have $P(X_1 \geq \delta) \geq 1/n$, for large enough n.

3. Let $\mathcal{P}$ be the set of all distributions on the real line with finite first moment. Show that there does not exist a function $f(x)$ such that $Ef(X) = \mu^2$ for all $P \in \mathcal{P}$ where μ is the mean of P, and X is a random variable with distribution P. We must have $h(x)dP(x) = \mu^2$ for all distributions on the real line with mean μ. If P is degenerate at a point y, this implies that $h(y) = y^2$ for all y. But if P has mean zero ($\mu = 0$) and is not degenerate, then $h(x)dP(x) = x^2dP(x) > 0 = \mu^2$. which is a contradiction.

4. Let g_1 and g_2 be estimable parameters within $\mathcal{P}$ with respective degrees m_1 and m_2.

   (a) Show $g_1 + g_2$ is an estimable parameter with degree $\leq \max(m_1, m_2)$.

   (b) Show $g_1 g_2$ is an estimable parameter with degree at most $m_1 + m_2$.

5. Look at the seminal paper "Probability Inequalities for Sums of Bounded Random Variables" by Wassily Hoeffding. It should be available via `lib.utexas.edu`. You can assume that n is a multiple of m (the degree of the kernel). Assume that the kernel is bounded, i.e. $|h(X_1, \ldots, X_m) - \theta| \leq b$, where $\theta = E[h(X_1, \ldots, X_m)]$.

   (a) Read and reproduce the proof of equation 5.7 for large sample deviation of order m U statistics.

   (b) Also prove Bernstein's inequality (see below) for U statistics. This is buried in the paper, you will have to find the bits and pieces and put them together. The Bernstein inequality is given by:

$$P(|U_n - \theta| \geq \epsilon) \leq a \exp\left(-\frac{n\epsilon^2/m}{c_1\sigma^2 + c_2\epsilon}\right),$$

   where $\sigma^2 = \mathrm{var}(h(X_1, \ldots, X_m))$ and a, c_1, c_2 are universal constants.