Homework Assignment 4

Due via Canvas, March 28th by midnight

SDS 384-11 Theoretical Statistics

1. Consider an i.i.d. sample of size n from a discrete distribution parametrized by $p_1, \ldots, p_{m-1}$ on m atoms. A common test for uniformity of the distribution is to look at the fraction of pairs that collide, or are equal. Call this statistic U.

   (a) Is U a U statistic? When is it degenerate?

   (b) What is the variance of U? Please give the exact answer, without approximation.

   (c) For a hypothesis test, we will consider alternative distributions which have $p_i = \frac{1+a}{m}$ for half of the atoms in the distribution and $\frac{1-a}{m}$ for the other half ($0 \leq a \leq 1$), for some $a > 0$. Assume that there are an even number of atoms. (Hint: think of this as a multinomial distribution.)

      i. What are the mean and variance of this statistic under the null?

      ii. What are the mean and variance of this under the alternative?

      iii. What is the asymptotic distribution of U under the null hypothesis that $p_i = 1/m$? *Hint: you can use the fact that for* $X_1, \ldots, X_N \overset{i.i.d}{\sim}$ *multinomial*$(q_1, \ldots, q_k)$, $\sum_{i=1}^{k}(N_i - Nq_i)^2/Nq_i \xrightarrow{d} \chi^2_{k-1}$, *where* N_i *is the number of datapoints with value* i.

      iv. Under the alternative hypothesis,is it always the case that U has a limiting normal distribution? Can you give a sufficient condition on the number of atoms m so that this is true? *Hint: Your variance will have two parts, and when the first one (with* $1/n$ *dependence on* n*) dominates the second (with* $1/n^2$ *dependence on* n*), you have a normal convergence. Typically, if* m *is small, the first one will dominate, however, it is possible that* m *is very large, in so you need* n *to be sufficiently large for the first term to dominate the second.*

2. (7 pts) 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 $\sigma^2 = \text{var}(h(X_1, \ldots, X_m))$ and a, c_1, c_2 are universal constants.

3. Compute the VC dimension of the following function classes. You can take it as everything on or inside the shape is +ve.

   (a) Circles in $\mathbb{R}^2$

   (b) Axis aligned rectangles in $\mathbb{R}^2$

   (c) Axis aligned squares in $\mathbb{R}^2$