Homework Assignment 4

Due via Canvas, April 10th 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. In class, you upper bounded the Rademacher complexity of a function class. Now you will derive a lower bound.
   (a) For function classes $\mathcal{F}$ with function values in $[0, 1]$, prove that $E\|\hat{P}_n - P\|_{\mathcal{F}} \geq \frac{\mathbb{R}_{\mathcal{F}}}{2} - \sqrt{\frac{\log 2}{2n}}$. *Hint: may be it is easier to start from* $\mathbb{R}_{\mathcal{F}}$ *and show that* $\mathbb{R}_F \leq 2E\|\hat{P}_n - P\|_{\mathcal{F}} + \sqrt{\frac{2\log 2}{n}}$. *In order to do this, you would need to add and subtract* $E[f(X)]$ *and then use triangle inequality.*
   (b) Now prove that $\|P - \hat{P}_n\|_{\mathcal{F}} \geq E\|P - \hat{P}_n\|_{\mathcal{F}} - \epsilon$ with probability at least $1 - \exp(-cn\epsilon^2)$ for some constant c.
   (c) Recall the class of all subsets with finite size in $[0, 1]$? Prove that then Rademacher complexity of this class is at least $1/2$. What does this imply?

3. Compute the VC dimension of the following function classes. You can take it as everything on or inside the shape is +ve. You should provide a complete proof of your answer.

   (a) Circles in $\mathbb{R}^2$
   (b) Axis aligned squares in $\mathbb{R}^2$
   (c) The function class $\{1(\sin(\theta x) \geq 0) : \theta \in \mathbb{R}\}$ for $x \in \mathbb{R}$