

# Homework Assignment 4

Due via Canvas, May 7th by midnight

SDS 384-11 Theoretical Statistics

1. Consider an i.i.d. sample of size  $n$  from a discrete distribution parametrized by  $p_1, \dots, p_{m-1}$  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  $X_1, \dots, X_N \stackrel{i.i.d.}{\sim} \text{multinomial}(q_1, \dots, q_k)$ ,  $\sum_{i=1}^k (N_i - Nq_i)^2 / Nq_i \xrightarrow{d} \chi_{k-1}^2$ , 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. (VC dimension) 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  $\mathbb{R}^2$
3. We will find the covering number of ellipses in this problem. Given a collection of positive numbers  $\{\mu_j, j = 1 \dots d\}$ , consider the ellipse

$$\mathcal{E} = \{\theta \in \mathbb{R}^d : \sum_i \theta_i^2 / \mu_i^2 \leq 1\} \quad (1)$$

- (a) Show that

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq d \log(1/\epsilon) + \sum_{j=1}^d \log \mu_j$$

*Hint: you can use the fact that the volume of the Ellipse defined in Eq 1 is given by  $\prod_{i=1}^d \mu_i \times C_d$  where  $C_d$  is the volume of a unit sphere in  $d$  dimensions. Extra credit for proving this! All you have to do is a simple substitution!*

- (b) Now consider an infinite-dimensional ellipse, specified by the sequence  $\mu_j = j^{-2\beta}$  for some parameter  $\beta > 1/2$ . Show that

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq C \left( \frac{1}{\epsilon} \right)^{1/2\beta},$$

where  $\|\theta - \theta'\|_{\ell_2}^2 = \sum_{j=1}^{\infty} (\theta_j - \theta'_j)^2$  is the squared  $\ell_2$ -norm on the space of square summable sequences. *Hint: This is going to involve a truncation argument. Truncate at  $d$  dimension, and obtain a relationship of the original covering number with the covering number of the truncated ellipse. Use your earlier result for  $d$  and then optimize over  $d$ .*