Homework Assignment 3

Due March 29th by midnight.

SDS 384-11 Theoretical Statistics

1. We will use the Efron Stein inequality to obtain bounds of variances for separately convex functions whose partial derivatives exist. A separately convex function $f(x_1, \ldots, x_n)$ is a convex function of its i^{th} variable, when all else are held fixed.

   (a) Let $X_1, \ldots, X_n$ be independent random variables taking values in the interval $[0, 1]$ and let $f : [0, 1]^n \to R$ be a separately convex function whose partial derivatives exist. Then $f(X) := f(X_1, \ldots, X_n)$ satisfies

   $$\mathrm{var}(f(X)) \leq E[\|\nabla f(X)\|^2]$$

   Hint: Recall that var$(Z) \leq \sum_i E(Z - E_i Z)^2 \leq \sum_i E(Z - Z_i)^2$, *where* $E_i[Z] = E[Z|X_{1:i-1}, X_{i+1:n}]$. *Define* $Z_i = \inf_x f(X_{1:i-1}, x, X_{i+1:n})$ *and then use convexity of* f.

   (b) Let A be a $m \times n$ random matrix with independent entries $A_{ij} \in [0, 1]$. Let

   $$Z = \sqrt{\lambda_1(A^T A)} = \sqrt{\sup_{u \in R^n : \|u\|=1} u^T A^T A u} = \sup_{u \in R^n : \|u\|=1} \|Au\|$$

   Show that $\mathrm{var}(Z) \leq 1$.

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,

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

   where c is some universal constant.

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

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

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

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

      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. In class we proved McDiarmid's inequality for bounded random variables. But now we will look at extensions for unbounded R.V's. Take a look at "Concentration in unbounded metric spaces and algorithmic stability" by Aryeh Kontorovich, `https://arxiv.org/pdf/1309.1007.pdf`. Reproduce the proof of theorem 1. The steps of this proof is very similar to the martingale based inequalities we looked at in class.

4. 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.*