CSE 291 Homework 2

Winter 2020

This homework is due in class on Tuesday, March 3rd. Please make sure that your solutions are written either with legible handwriting or on LaTeX. You are welcome to discuss problems with other students (though are encouraged to think about the problems on your own before looking for help), but are asked to do your solution writeups on your own. All problems require full justification of your work.

**Question 1** (VC-Dimensions of Unions, 20 points). Let \( A \) and \( B \) be families of sets each of VC-dimension at most \( d \). Let \( C \) be the family \( \{ A \cap B : A \in C, B \in B \} \). Prove that the VC-dimension of \( C \) is at most \( O(d) \).

**Hint:** Consider the sizes of \( S_C(n) \) for appropriately large \( n \).

**Question 2** (Learning Concave Distributions, 40 points). Define a concave distribution to be a continuous distribution supported on some interval \( I \subseteq \mathbb{R} \) where the probability density function \( p(x) \) on \( I \) is concave.

- Show that any such distribution can be \( \epsilon \)-approximated in total variational distance by an \( O(\epsilon^{-1/2}) \)-piecewise linear distribution. **Hint:** If \( I = [0, 1] \), and the derivative is bounded take \( O(\epsilon^{-1/2}) \) points along the graph of \( p(x) \) and use their convex hull as the graph of the approximating distribution. Place a point every time the value of \( x \) or \( p'(x) \) changes by more than \( \epsilon^{-1/2} \). When \( p' \) is not bounded, do a rotated version of this on the sides of the graph. [Note: This implies that we can learn such distributions with \( O(\epsilon^{-5/2}) \) samples.]

- Show that learning a concave distribution to error \( \epsilon \) requires \( \Omega(\epsilon^{-5/2}) \) samples. **Hint:** Begin by considering some fixed distribution (for example \( p(x) = 4/3 - x^2 \) on \( [0, 1] \)), split it into \( \epsilon^{-1/2} \) regions and modify \( p \) slightly on each region to keep it concave but change the distribution by \( \epsilon \) in total variational distance.

**Note:** These results can also be generalized to the much more standard example of log-concave distributions.

**Question 3** (\( A_k \)-Closeness Testing Lower Bounds, 40 points). To prove lower bounds for \( A_k \)-closeness testing, we begin by assuming that the tester takes \( \text{Poi}(m) \) samples from each of \( p \) and \( q \) and then returns an answer based on the sorted order of the elements (one can show that a general tester can be reduced to this form). Your hard distribution will need to look something like the following. The line should be split into about \( m + k \) segments. About \( m \) of those segments should have \( p = q \) each with total mass about \( 1/m \). The other segments should either consist of a block where \( p = q \) each of mass \( \epsilon/k \) or an interval of mass \( \epsilon/(2k) \) of \( p \) a block of mass \( \epsilon/k \) of \( q \) followed by another block of mass \( \epsilon/(2k) \) of \( p \).

Note that in the first case, \( p = q \) whereas in the latter, the \( A_k \) distance is \( \Omega(\epsilon) \). Also note that if any two samples are taken from a given block, their order provides no information about which case we are in. Thus all information to distinguish these cases must come from blocks with at least 3 samples.

Show along these lines that for \( m \leq k \) and \( m = o(k^{4/5}/\epsilon^{6/5}) \) that there is no sample-order-based \( A_k \)-closeness tester with \( m \) samples.