Homework 9 Solutions

Problem 1.

(a) Consider any target concept $[a, b]$ where $0 < a < b < 1$. Consider a distribution which places probability weight of $1/4$ on each of four points a^-, a^+, b^-, b^+ where $0 < a^- < a < a^+ < b^- < b < b^+ < 1$. In order to have an error rate less than $1/4$, a statistical query learning algorithm must output an hypothesis $[a', b']$ where $a^- < a' \leq a^+$ and $b^- \leq b' < b^+$. Note that an adversary can choose a^- and a^+ (respectively, b^- and b^+) *arbitrarily* close to a (respectively, b). At this point, it is not hard to see that a statistical query learning algorithm will require an unbounded number of $STAT$ queries. One could argue this formally by showing that each $STAT$ query imparts at most one new bit of accuracy for either a or b (e.g. by standard binary search), but that as a^- and a^+ (respectively, b^- and b^+) approach a (respectively, b), the number of bits of accuracy required can be arbitrarily high.

(b) We first describe a method of "discretizing" the interval $[0, 1]$. Our goal will be to break up the $[0, 1]$ interval into subintervals whose distribution weights are at most ϵ with probability at least $1 - \delta$. We will then use this procedure (with different ϵ and δ parameters) to aid in learning $\mathcal{C}_I$ in the statistical query model with unlabelled examples.

For the purposes of analysis, consider dividing the interval $[0, 1]$ into $2/\epsilon$ consecutive subintervals of *distribution weight* $\epsilon/2$. (Note: If the distribution contains point-masses of significant weight, this procedure will have to be modified slightly. The idea is the same, however.) Now, we would like to draw m examples from $UEX(D)$ such that with probability at least $1-\delta$, every subinterval will be hit by at least one example. We determine m as follows:

$$\begin{aligned} \Pr\{\text{A particular subinterval is missed by all } m \text{ examples}\} &\leq (1 - \epsilon/2)^m \\ \Pr\{\text{Some subinterval is missed by all } m \text{ examples}\} &\leq \tfrac{2}{\epsilon}(1 - \epsilon/2)^m \end{aligned}$$

Bounding this quantity by δ, we find that choosing

$$m = \tfrac{2}{\epsilon} \ln \tfrac{2}{\epsilon\delta}$$

is sufficient. Now, sort the examples drawn (including the endpoints 0 and 1) and consider the weight of any region between two consecutive examples. If all $2/\epsilon$ consecutive subintervals of distribution weight $\epsilon/2$ are hit by some example, then the weight of any region between two consecutive examples must be less than ϵ. (If the weight between two consecutive examples is at least ϵ, then there must be a subinterval of weight $\epsilon/2$ contained therein which was missed.)

Run the above discretization procedure with parameters $\epsilon' = \epsilon/2$ and $\delta' = \delta/2$. We choose $\epsilon/2$ in order to obtain sufficiently accurate estimates of both a and b. We choose $\delta/2$ in order to retain a $\delta/2$ failure probabilty for the simulation of the $STAT$ oracle. Sort the m' examples drawn and label them $x_1, x_2, \ldots, x_{m'}$. Ask queries of the form, "What is the probability weight of positive examples in the range $[0, x_i]$ to within $\epsilon/4$?" Repeat queries for $i = 1, 2, \ldots$ until the first point at which the value returned is greater than $\epsilon/4$; let this value of i be i_l. This will not occur until after $[0, x_{i_l}]$ contains *some* positive examples. Further, $[0, x_{i_l - 1}]$ will contain *at most* $\epsilon/2$ weight of positive examples. Finally, since $[x_{i_l - 1}, x_{i_l}]$ has weight at most $\epsilon/2$, at most $\epsilon/2$ weight of negative examples will be contained in this range. Thus, whether $x_{i_l - 1}$ is to the left or right of the true left endpoint, at most $\epsilon/2$ probability weight of examples will be misclassified by this choice of a left endpoint. Let our hypothesis' left endpoint be $x_{i_l - 1}$. Repeat in a similar manner to find the right endpoint. The total error rate of the ouput hypothesis will be at most ϵ as required.

(c) In class, we showed that efficient learnability in the statistical query model implies both efficient PAC-learnability and efficient PAC-learnability in the presence of classification noise. The proof for efficient learnability in the statistical query model with unlabelled examples is identical except that we need to simulate the new $UEX(D)$ oracle. This can be done trivially, however, by drawing examples from either $EX(c, D)$ or $EX^\eta(c, D)$ and simply removing the labels.