A. PAC learning

1. Give a PAC-learning algorithm for the concept class $C$ formed by closed intervals $[a, b]$ with $a, b \in \mathbb{R}$. You should give a careful proof as for the case of axis-aligned rectangles described in class.

Solution: Given a sample $S$, one algorithm consists of returning the tightest closed interval $I_S$ containing positive points. Let $I = [a, b]$ be the target concept. If $\Pr[I] < \epsilon$, then clearly $R(I_S) < \epsilon$. Assume that $\Pr[I] \geq \epsilon$. Consider two intervals $I_L$ and $I_R$ defined as follows:

$I_L = [a, x]$ with $x = \inf\{x : \Pr[a, x] \geq \epsilon/2\}$
$I_R = [x', b]$ with $x' = \sup\{x' : \Pr[x', b] \geq \epsilon/2\}$.

By the definition of $x$, the probability of $[a, x]$ is less than or equal to $\epsilon/2$, similarly the probability of $]x', b]$ is less than or equal to $\epsilon/2$. Thus, if $I_S$ overlaps both with $I_L$ and $I_R$, then its error region has probability at most $\epsilon$. Thus, $R(I_S) > \epsilon$ implies that $I_S$ does not overlap with either $I_L$ or $I_R$, that is either none of the training points falls in $I_L$ or none falls in $I_R$. Thus, by the union bound,

$$\Pr[R(I_S) > \epsilon] \leq \Pr[S \cap I_L = \emptyset] + \Pr[S \cap I_R = \emptyset] \leq 2(1 - \epsilon/2)^m \leq 2e^{-m\epsilon/2}.$$ 

Setting $\delta$ to match the right-hand side gives the sample complexity $m = \frac{2}{\epsilon} \log \frac{2}{\delta}$ and proves the PAC-learning of closed intervals. \qed

2. Give a PAC-learning algorithm for the concept class $C_2$ formed by unions of two closed intervals, that is $[a, b] \cup [c, d]$, with $a, b, c, d \in \mathbb{R}$. Extend your result to derive a PAC-learning algorithm for the concept class $C_p$ formed by unions of $p \geq 1$ closed intervals, thus $[a_1, b_1] \cup \cdots \cup [a_p, b_p]$, with $a_k, b_k \in \mathbb{R}$ for $k \in [1, p]$. What are the time and sample complexities of your algorithm as a function of $p$?
Solution: Given a sample $S$, our algorithm consists of the following steps:

(a) Sort $S$ in ascending order.
(b) Loop through sorted $S$, marking where intervals of consecutive positively labeled points begin and end.
(c) Return the union of intervals found on the previous step. This union is represented by a list of tuples that indicate start and end points of the intervals.

This algorithm works both for $p = 2$ and for a general $p$. We will now consider the problem for $C_2$. To show that this is a PAC-learning algorithm, we need to distinguish between two cases.

The first case is when our target concept is a disjoint union of two closed intervals: $I = [a, b] \cup [c, d]$. Note that unlike the first part of the question, there are two sources of error: false negatives in $[a, b]$ and $[c, d]$ and also false positives in $(b, c)$. False positives may occur if no sample is drawn from $(b, c)$. By linearity of expectation and since these two error regions are disjoint, we can write

$$R(h_S) = R_{FP}(h_S) + R_{FN,1}(h_S) + R_{FN,2}(h_S),$$

where

$$R_{FP}(h_S) = \Pr_{x \sim D}[x \in h_S, x \notin I],$$

$$R_{FN,1}(h_S) = \Pr_{x \sim D}[x \notin h_S, x \in [a, b]],$$

$$R_{FN,2}(h_S) = \Pr_{x \sim D}[x \notin h_S, x \in [c, d]].$$

Since we need to have that at least one of $R_{FP}(h_S)$, $R_{FN,1}(h_S)$, $R_{FN,2}(h_S)$ exceeds $\epsilon/3$ in order for $R(h_S) > \epsilon$, by the union bound, the following holds

$$\Pr(R(h_S) > \epsilon) \leq \Pr((R_{FP}(h_S) > \epsilon/3) \lor (R_{FN(h_S),1} > \epsilon/3) \lor (R_{FN(h_S),2} > \epsilon/3))$$

$$\leq \Pr(R_{FP}(h_S) > \epsilon/3) + \sum_{i=1}^{2} \Pr(R_{FN(h_S),i} > \epsilon/3). \quad (1)$$

We first bound $\Pr(R_{FP}(h_S) > \epsilon/3)$. Note that if $R_{FP}(h_S) > \epsilon/3$, then $\Pr((b, c)) > \epsilon/3$ and hence

$$\Pr(R_{FP}(h_S) > \epsilon/3) \leq (1 - \epsilon/3)^m \leq e^{-mc/3}.$$
Now we can bound $\Pr(R_{\text{FN}(h_S),i} > \epsilon/3)$ by $2e^{-m\epsilon/6}$ using the same argument as in the previous question. Therefore,

$$\Pr(R(h_S) > \epsilon) \leq e^{-m\epsilon/3} + 4e^{-m\epsilon/6} \leq 5e^{-m\epsilon/6}.$$ 

Setting the right-hand side to $\delta$ and solving for $m$ yields the condition $m \geq \frac{6}{\epsilon} \log \frac{5}{\delta}$.

The second case that we need to consider is when $I = [a,d]$, that is $[a,b] \cap [c,d] \neq \emptyset$. In that case, our algorithm reduces to that of the previous question and it was already shown that $m \geq \frac{2}{\epsilon} \log \frac{2}{\delta}$ samples suffice to learn this concept. Therefore, we conclude that our algorithm is indeed a PAC-learning algorithm.

The extension of this result to the case of $C_p$ is straightforward. The only difference is that in (1), one has two summations for $p-1$ regions of false positives and $2p$ regions of false negatives. In that case, the sample complexity is $m \geq \frac{2(2p-1)}{\epsilon} \log \frac{3p-1}{\delta}$.

The sorting step of our algorithm takes $O(m \log m)$ time and the steps (b) and (c) are linear in $m$, which leads to an overall time complexity of $O(m \log m)$.

**B. Rademacher complexity, growth function**

1. Let $H$ be the family of threshold functions over the real line: $H = \{x \mapsto 1_{x \leq \theta}: \theta \in \mathbb{R}\} \cup \{x \mapsto 1_{x > \theta}: \theta \in \mathbb{R}\}$. Give an upper bound on the growth function $\Pi_m(H)$. Use that to derive an upper bound on $\mathcal{R}_m(H)$.

   **Solution:** Given $m$ distinct points on the line, there are at most $m+1$ choices of the threshold $\theta$ leading to distinct classifications: between two points or before/after all points. Since there are two choices (classifying those to right as positive or those to the left), there are $2(m+1)$ choices. Thus, $\Pi_m(H) \leq 2m$ and, by the bound on the Rademacher complexity shown in class, $\mathcal{R}_m(H) \leq \sqrt{\frac{2\log(2m)}{m}}$.

2. Let $H_1$ and $H_2$ be two families of functions mapping $\mathcal{X}$ to $\{0, 1\}$ and let $H = \{h_1h_2: h_1 \in H_1, h_2 \in H_2\}$. Show that the empirical Rademacher complexity of $H$ for any sample $S$ of size $m$ can be bounded as follows:

   $$\hat{\mathcal{R}}_S(H) \leq \hat{\mathcal{R}}_S(H_1) + \hat{\mathcal{R}}_S(H_2).$$
**Hint:** use the Lipschitz function $x \mapsto \max(0, x - 1)$ and Talagrand’s contraction lemma (see textbook or slides).

Use that to bound the Rademacher complexity $\mathcal{R}_m(U)$ of the family $U$ of intersections of two concepts $c_1$ and $c_2$ with $c_1 \in C_1$ and $c_2 \in C_2$ in terms of the Rademacher complexities of $C_1$ and $C_2$.

**Solution:** This is a recent result due to [1] (see that reference and other related publications for other results of this type).

Observe that for any $h_1 \in H_1$ and $h_2 \in H_2$, we can write $h_1 h_2 = (h_1 + h_2 - 1)1_{h_1 + h_2 - 1 \geq 0} = (h_1 + h_2 - 1)_+$. Since $x \mapsto (x - 1)_+$ is 1-Lipschitz over $[0, 2]$, by Talagrand’s contraction lemma, the following holds: $\hat{\mathcal{R}}_S(H) \leq \hat{\mathcal{R}}_S(H_1 + H_2) \leq \hat{\mathcal{R}}_S(H_1) + \hat{\mathcal{R}}_S(H_2)$.

Since concepts can be identified with indicator functions, the intersection of two concepts can be identified with the product of two such indicator functions. In view of that, by the result just proven and after taking expectations, the following holds:

$$\mathcal{R}_m(C) \leq \mathcal{R}_m(C_1) + \mathcal{R}_m(C_2).$$

**References**