A. Learning union of intervals

1. Let \([a, b]\) and \([c, d]\) be two intervals of the real line with \(a \leq b \leq c \leq d\). Let \(\epsilon > 0\) and assume that \(\Pr_D((b, c)) > \epsilon\), where \(D\) is the distribution according to which points are drawn. Show that the probability that \(m\) points be drawn i.i.d. without any of them falling in the interval \((b, c)\) is at most \(e^{-m\epsilon}\).

The probability of each point falling outside the interval is at most \((1 - \epsilon)\), thus the probably of the event is bounded by \((1 - \epsilon)^m \leq e^{-m\epsilon}\).

2. Show that the concept class formed by unions of two closed intervals in \(\mathbb{R}\), e.g., \([a, b] \cup [c, d]\), is PAC-learnable by giving a proof similar to the one given in Lecture 2 for axis-aligned rectangles (hint: your algorithm might not return a hypothesis consistent with future negative points in this case).

One simple PAC-learning algorithm can be defined as follows. The algorithm returns the following hypothesis \(h_S\) for a training sample \(S\):

(a) if there are two separate sequences of positively labeled points in the training data (separated by negative points), the return the union of two intervals \([a', b'] \cup [c', d']\) with \([a', b'] \subseteq [a, b]\) and \([c', d'] \subseteq [c, d]\), where \([a', b']\) is the smallest interval containing the first sequence of positive points and \([c', d']\) the smallest one containing the second sequence of positive points;

(b) otherwise, return the smallest interval \([a', d']\) containing all the positive points, which can be written as the union of two closed intervals.

Let \([a, b] \cup [c, d]\) be the target concept. Let \(\epsilon > 0\). We can assume that \(\Pr[[a, b]] > \epsilon/3\) and \(\Pr[[c, d]] > \epsilon/3\). Other cases are either trivial or simple to analyze as for what follows. As in the proof for axis-aligned rectangles, consider four regions \(r_1, r_2, r_3\) and \(r_4\) defined as follows. \(r_1\) is an interval of the form \([a, b']\), \(b' \leq b\) such that \(\Pr[[a, b']] = \epsilon/6\). Similarly, \(r_2, r_3\) and \(r_4\) are regions bordering the endpoints of the two intervals, each with probability \(\epsilon/6\).
Now, by the definition of the algorithm and a geometric argument similar to the case of axis-aligned rectangles, if $error(h_S) > \epsilon$, then either the union of intervals predicted misses at least one of the regions $r_i$, $i \in [1, 4]$, or $Pr[(b, c)] > \epsilon/3$ and no training point falls in $(b, c)$ (second case of the hypothesis returned by the algorithm). Thus, by the union bound and in view of previous question:

$$Pr[error(h_S) > \epsilon] \leq Pr[\exists i \in [1, 4]: h_s \text{ misses } r_i] + e^{-me/3}$$

$$\leq \sum_{i=1}^{4} Pr[h_s \text{ misses } r_i] + e^{-me/3}$$

$$\leq 4(1 - \epsilon/6)^m + e^{-me/3} \leq 4e^{-me/6} + e^{-me/3} \leq 5e^{-me/6}.$$  

Setting $\delta > 0$ to match the upper bound yields that for $m \geq \frac{6}{\epsilon} \log \frac{5}{\delta}$, with probability at least $1 - \delta$, $error(h_s) \leq \epsilon$.

B. Consistent hypotheses

- In Lecture 2, we showed that for a finite hypothesis set $H$, a consistent learning algorithm $L$ is a PAC-learning algorithm. Here, we consider a converse question. Let $Z$ be a finite set of $m$ labeled points. Suppose that you are given a PAC-learning algorithm $L$. Show that you can use $L$ to find in polynomial time a hypothesis $h \in H$ that is consistent with $Z$, with high probability (hint: you can select an appropriate distribution $D$ and give a condition on $error(h)$ for $h$ to be consistent).

Since PAC-learning with $L$ is possible for any distribution, let $D$ be the uniform distribution over $Z$. Note that, in that case, the cost of an error of a hypothesis $h$ on any point $z \in Z$ is $Pr_D[z] = 1/m$. Thus, if $error_D(h) < 1/m$, we must have $error_D(h) = 0$ and $h$ is consistent. Thus, choose $\epsilon = 1/(m+1)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $S$ with $|S| \geq P((m+1), 1/\delta)$ points (where $P$ is some fixed polynomial) the hypothesis $h_S$ returned by $L$ is consistent with $Z$ since $error_D(h_S) \leq 1/(m + 1)$.

C. Rademacher complexity

- Professor Jesetoo claims to have found a better bound on the Rademacher complexity of any hypothesis set $H$ of functions taking values in $\{-1, +1\}$, in terms of its VC dimension $VCdim(H)$. His bound is of the form $\mathbb{R}_m(H) \leq O(\frac{VCdim(H)}{m})$. Can you show that Professor Jesetoo’s claim cannot be correct? (hint: consider a hypothesis set $H$ reduced to just two simple functions.)
Consider the simple case where $H$ is reduced to the constant hypothesis $h_1: x \mapsto 1$ and $h_{-1}: x \mapsto -1$. Then, by definition of the empirical Rademacher complexity,

$$\hat{\mathcal{R}}_S(H) = \frac{1}{m} E[\max\{\sum_{i=1}^{m} \sigma_i, \sum_{i=1}^{m} -\sigma_i\}] = \frac{1}{m} E[\|\sum_{i=1}^{m} \sigma_i\|]$$

Let $X = \sum_{i=1}^{m} \sigma_i$. Note that $E[X^2] = E[\sum_{i,j=1}^{m} \sigma_i \sigma_j]$. For any $i \neq j$, since $\sigma_i$ and $\sigma_j$ are independent, $E[\sigma_i \sigma_j] = E[\sigma_i] E[\sigma_j] = 0$. Thus,

$$E[X^2] = \sum_{i=1}^{m} E[\sigma_i \sigma_i] = \sum_{i=1}^{m} E[\sigma_i^2] = m.$$ 

Now, by Hölder’s inequality,

$$m = E[X^2] = E[|X|^{2/3} |X|^{4/3}] \leq E[|X|^{2/3}] E[|X|^{4/3}].$$

Thus,

$$E[|X|] \geq \frac{m^{3/2}}{E[|X|]^{1/2}} = \frac{m^{3/2}}{\sqrt{E[\sum_{i=1}^{m} \sigma_i^4 + 3 \sum_{i \neq j} \sigma_i^2 \sigma_j^2]}} = \frac{m^{3/2}}{\sqrt{m + 3m(m-1)}} = \frac{m^{3/2}}{\sqrt{m(3m-2)}} \geq \frac{m^{3/2}}{\sqrt{m(3m)} = \frac{\sqrt{m}}{\sqrt{3}}.}

$$

This shows that

$$\hat{\mathcal{R}}_S(H) \geq \frac{\sqrt{m}}{\sqrt{3}}.$$ 

Since $\mathcal{R}_m(H) \geq \hat{\mathcal{R}}_S(H) + O(\frac{1}{\sqrt{m}})$, it implies $\mathcal{R}_m(H) \geq O(\frac{1}{\sqrt{m}})$, which contradicts $\mathcal{R}_m(H) \leq O(\frac{1}{m})$.

Note that for the lower bound, we could have used instead a more general result (Khintchine’s inequality) which states that for any $a \in \mathbb{R}^m$,

$$E[\|\sigma \cdot a\|] \geq \|a\|_2 \sqrt{2}.$$