1. (a) We want a bound on

$$P\left(L_D(A(S)) \geq \frac{1}{8}\right)$$

Markov's inequality directly gives us an upper bound, but we want a <u>lower</u> bound

Since our NFL theorem is for binary classification with the 0-1 loss, we know that

$$L_D(A(S)) \in [0,1]$$

Now consider

$$P\left(L_D(A(S)) \geq \frac{1}{8}\right)$$

$$= P\left(1 - L_D(A(S)) \leq \frac{7}{8}\right)$$

To lower bound this, we can upper bound

$$P\left(1 - L_D(A(S)) \geq \frac{7}{8}\right)$$

also $\in [0,1]$

which is $\leq \dfrac{\mathbb{E}\left[1 - L_D(A(S))\right]}{7/8}$ (Markov's Inequality)

$$= \frac{1 - \mathbb{E}\left[L_D(A(S))\right]}{7/8}$$

$$\leq \frac{1 - 1/4}{7/8} \quad \text{(from given bound on exp.)}$$

$$= \frac{8-2}{7} = \frac{6}{7}$$

$$\Rightarrow P\left(1 - L_D(A(S)) \geq \frac{7}{8}\right) \leq \frac{6}{7}$$

$$P\left(1 - L_D(A(S)) \leq \frac{7}{8}\right) \geq 1 - \frac{6}{7} = \frac{1}{7}$$

$\Rightarrow P\left(L_D(A(S)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$.

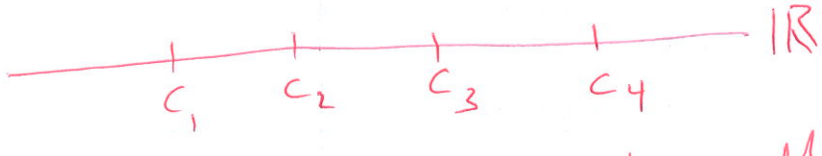
(b) The proof involves selecting a set $C \subset \mathcal{X}$ of size $2N$ for <u>any</u> given training set of size N (and the condition that $N < \frac{|\mathcal{X}|}{2}$ ensures this can be done). So the choice of N is <u>prior</u> to the selection of C, and if you demand to see more data (subject to $N < \frac{|\mathcal{X}|}{2}$), I can always correspondingly choose a bigger C, and the proof of non-learnability still goes through.

2. We start with the guess that

$$VC\dim(\mathcal{H}) = 4$$

as it has 4 parameters.

i) Consider a set $C = \{c_1, c_2, c_3, c_4\}$



To show shattering, consider all possible labellings ($2^4 = 16$). Break them up by no. of points labelled 1.

a) 0 points labelled 1: can choose $a \leq b < c_1$; $d \geq c > c_4$

b) 1 point labelled 1, say c_i : choose

$$\underbrace{c_{i-1} <}_{\text{if exists}} a \le c_i \le b \underbrace{< c_{i+1}}_{\text{if exists}} \quad ; \quad d \ge c > c_4$$

c) 2 points : Say c_i, c_j : choose ($c_j > c_i$)

$$\overbrace{c_{i-1} <}^{\text{if exists}} a \le c_i \le b < c_{i+1}$$

$$\underbrace{c_{j-1} <}_{\cancel{\text{if exists}}} c \le c_j \le d \underbrace{< c_{j+1}}_{\text{if exists}}$$

d) 3 points : Say c_i, c_j, c_k ($c_i < c_j < c_k$)

At least two of them must be successive points : either c_i, c_j or c_j, c_k

If c_i, c_j successive, choose :

$$\overbrace{c_{i-1} <}^{\text{if exists}} a \le c_i \; ; \; c_j \le b < c_{j+1}$$

$$c_{k-1} < c \le c_k \le d \underbrace{< c_{k+1}}_{\text{if exists}}$$

Else, choose :

$$\overbrace{c_{i-1} <}^{\text{if exists}} a \le c_i \le b < c_{i+1}$$

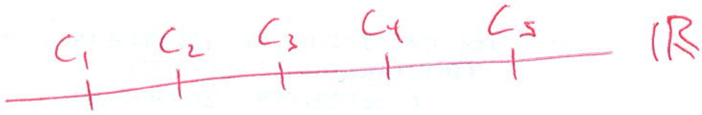
$$c_{j-1} < c \le c_j \; ; \; c_k \le d \underbrace{< c_{k+1}}_{\text{if exists}}$$

e) 4 points : Choose $a \le c_1$, $b \ge c_4$; $d \ge c \ge b$

Hence all possible labellings achievable by $\mathcal{H}$.

$\Rightarrow \mathcal{H}$ shatters C.

ii) Consider any set C of size 5. Without loss of generality, $C = \{c_1, c_2, c_3, c_4, c_5\}$ where $c_1 < c_2 < c_3 < c_4 < c_5$. Consider the labelling:



Label: 1 0 1 0 1

With only 2 positive intervals we cannot achieve 1 labels in 3 discontiguous parts of the input space. More specifically, we need

$$a \leq c_1 \leq b < c_2 \quad \text{(to achieve correct labels for } c_1, c_2\text{)}$$

then

$$c_2 < c \leq c_3 \leq d < c_4$$

(to also correctly label c_3, c_4)

but now, $c_5 > c_4 > d$, so c_5 cannot be labelled as 1.

Hence sets of size 5 cannot be shattered by $\mathcal{H}$.

From i) and ii), $\underline{VC\dim(\mathcal{H}) = 4}$.

3. (a) The key issue is the implicit assumption that $N_{\mathcal{H}}(\epsilon, \delta)$, as defined, is finite. The maximum of an <u>infinite set</u> of finite quantities <u>need not</u> be finite.

For example, every number in $\mathbb{N}$ is finite; but $\max(\mathbb{N})$ is not finite! Similarly, relating to the polynomials example, the degree of polynomials in each $\mathcal{H}_n$ was finite; but the max. degree over the union of all $\mathcal{H}_n$ was not. So, in case of infinite $\mathcal{H}$, being able to compete with each individual hypothesis does <u>not</u> guarantee being able to compete with the best one, because finding out what the latter is may require infinite data.

(b) The intuition arises from implicitly thinking of maximisation over <u>finite</u> sets. So clearly, the argument is valid whenever $\mathcal{H}$ is indeed finite (and we already know that all such $\mathcal{H}$ are PAC learnable). More generally, the argument would hold precisely when $N_{\mathcal{H}}(\epsilon, \delta)$, as defined here, happens to be finite; or the set $\{N_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) : h \in \mathcal{H}\}$ happens to be a <u>bounded</u> set $\forall \epsilon, \delta$. From the Fundamental Theorem, we can say that this will happen iff $VCdim(\mathcal{H}) < \infty$.