1. (a) - Specification of conditions which lead to valid, generalisable learning, as opposed to overfitting.
- Principles or mechanisms to developed which help machine learning algorithms to filter out the wheat (useful learning) from the chaff (overfitting).

(b) Inductive bias or prior knowledge plays a key role in making learning useful, and avoiding overfitting. So formally specifying inductive bias, and its role in learnability, is a key goal of learning theory.

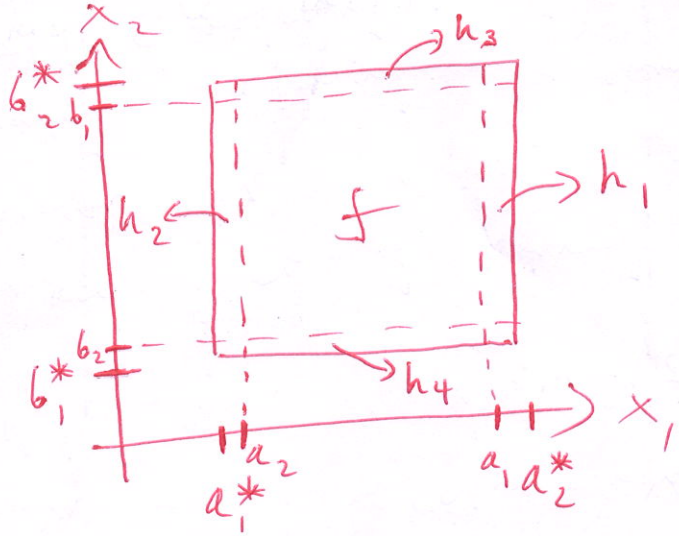
2. (a) If we let $h_S = ERM_{\mathcal{H}}(S)$, then by realisability, we must have $L_S(h_S) = 0$. If this property holds for A, then it will be an ERM learner on $\mathcal{H}$.

Since A returns the smallest rectangle enclosing all positive instances, it cannot contain negative instances: otherwise there would have to be a smaller rectangle which successfully separates the classes, given realisability.

So A returns a rectangle containing all positives and no negatives: hence its loss on S is always 0, and it is an ERM.

(b) i. Similar argument to part (a): no tighter rectangle than that returned by A can separate the data, by definition. Since f of course must separate the data, it must fully contain the rectangle returned by A.

ii.



$h_S = A(S)$, so it is the tightest rectangle containing all positives in S.

If S contains at least one positive from each of $h_1 - h_4$ above, then clearly the rectangle returned by A must contain within it the rectangle $h_{(a_2, a_1, b_2, b_1)}$. Note that the true loss of h_S can only come from instances contained in between it and f, as h_S is entirely inside of f (part i).

The total probability of observing an instance from this region is upper bounded by the combined prob. mass of the area of $h_1 - h_4$, which is $\leq 4 \times \frac{\epsilon}{4} = \epsilon$.

Hence we have:

$$L_{\mathcal{D},f}(h_S) \leq \epsilon.$$

iii. We want to bound, $\forall i \in \{1,2,3,4\}$,

$$\mathcal{D}^N\left(\{S|_{\underline{x}} : \forall n \in \{1,..N\}, h_i(x_n) = 0\}\right)$$

Since S is sampled i.i.d., this is

$$= \prod_{n=1}^{N} \mathcal{D}\left(\{x_n : h_i(x_n) = 0\}\right)$$

Since the prob. mass of h_i is $\frac{\epsilon}{4}$,

$$= \prod_{n=1}^{N} \left(1 - \frac{\epsilon}{4}\right) = \left(1 - \frac{\epsilon}{4}\right)^N.$$

iv. We want a union bound on the prob. that S does not contain instances of tasty pitayas covering all of $h_1 - h_4$.

This prob. is

$$\mathcal{D}^N\left(\{S|_{\underline{x}} : \exists i \in \{1,2,3,4\} : \forall n \in \{1,..N\}, h_i(x_n) = 0\}\right)$$

$$= \mathcal{D}^N\left(\bigcup_{i=1}^{4} \{S|_{\underline{x}} : \forall n \in \{1,..N\}, h_i(\underline{x}_n) = 0\}\right)$$

$$\left(\begin{matrix}\text{Union}\\ \text{Bound}\end{matrix}\right) \leq \sum_{i=1}^{4} \mathcal{D}^N\left(\{S|_{\underline{x}} : \forall n \in \{1,..N\}, h_i(\underline{x}_n) = 0\}\right)$$

$$\left(\begin{matrix}\text{part}\\ \text{iii}\end{matrix}\right) \leq \sum_{i=1}^{4} \left(1-\frac{\epsilon}{4}\right)^N = \underline{4\left(1-\frac{\epsilon}{4}\right)^N}.$$

v. For PAC to hold, we need

$$\mathcal{D}^N\left(\{S|_{\underline{x}} : L_{\mathcal{D},f}(h_S) > \epsilon\}\right) \leq \delta$$

We have shown that $L_{\mathcal{D},f}(h_S) \leq \epsilon$ whenever S has tasty papayas in each of h_1-h_4; and that the prob. of this not being true is at most bounded by $4\left(1-\frac{\epsilon}{4}\right)^N$. Hence:

$$\mathcal{D}^N\left(\{S|_{\underline{x}} : L_{\mathcal{D},f}(h_S) > \epsilon\}\right) \leq 4\left(1-\frac{\epsilon}{4}\right)^N$$

So we need

$$4\left(1-\frac{\epsilon}{4}\right)^N \leq \delta$$

$$\left(1-\frac{\epsilon}{4}\right)^N \leq \frac{\delta}{4}$$

$$N \log\left(1-\frac{\epsilon}{4}\right) \leq \log\frac{\delta}{4}$$

Using $1-\epsilon \leq e^{-\epsilon}$:

$$N \log\left(1-\frac{\epsilon}{4}\right) \leq N \log\left(e^{-\epsilon/4}\right)$$

$$= \frac{-N\epsilon}{4}$$

So we need

$$\frac{-N\epsilon}{4} \leq \log\frac{\delta}{4}$$

$$\frac{N\epsilon}{4} \geq \log\frac{4}{\delta}$$

$$\boxed{N \geq \frac{4\log(4/\delta)}{\epsilon}}$$

Thus, we have:

$$N_H(\epsilon,\delta) \leq \frac{4\log(4/\delta)}{\epsilon}.$$