1. (a) For any function $f : \mathcal{X} \to \{0,1\}$ we can construct a corresponding distrubution

$$\mathcal{P} = \begin{cases} \frac{1}{|\mathcal{X}|} & (x,y) : y = f(x) \\ 0 & \text{otherwise} \end{cases}$$

for which $L_{\mathcal{P}}(f) = 0$

(b) There are $T = 2^{|\mathcal{X}|}$ possible functions $f_i : \mathcal{X} \to \{0,1\}$ for which we can define corresponding disturbutions $\mathcal{P}_i$ as above. Assume we draw a training set $|C|$ from $\mathcal{X} \times \{0,1\}$ of $m \leq |\mathcal{X}|/2$ samples. We have an equal probability of drawing any of $k = |\mathcal{X}|^m$ sequences from $\mathcal{X}$. Denote the family of such sequences by $S_1, \ldots, S_k$ and let S_j^i denote the j^{th} sequence of tuples corresponding to f_i such that $S_j^i = (x_{j_1}, f(x_{j_1}), \ldots, (x_{j_m}, f(x_{j_m}))$.

With this setup, for a specific distribution $\mathcal{P}_i$ we have

$$\mathop{\mathbb{E}}_{S \sim \mathcal{P}_i^m}[L_{\mathcal{P}_i}(A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{P}_i}(A(S_j^i))$$

We can also exploit properties of the max$(\cdot)$ and min$(\cdot)$ functions to write

$$\begin{aligned}
\max_{i \in [T]} \frac{1}{k}\sum_{i=1}^{k} L_{\mathcal{P}_i}(A(S_j^i)) &\geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{P}_i}(A(S_j^i)) \\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{P}_i}(A(S_j^i)) \\
&\geq \min_{j \in [k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{P}_i}(A(S_j^i))
\end{aligned}$$

Now let $p = |\mathcal{X}| - m$ such that $p \geq m$, and let $v_1, \ldots, v_p$ represent the samples not included in S_j. Then

$$L_{\mathcal{P}_i}(h) = \frac{1}{|\mathcal{X}|}\sum_{x \in \mathcal{X}} \mathbb{1}_{[h(x) \neq f_i(x)]} \geq \frac{1}{2p}\sum_{r=1}^{p} \mathbb{1}_{[h(v_r) \neq f_i(v_r)]}$$

and

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{P}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}\mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \geq \frac{1}{2}\min_{r \in [p]}\frac{1}{T}\sum_{i=1}^{T}\mathbb{1}_{[h(v_r) \neq f_i(v_r)]}.$$

For any fixed r we can split $\{f_i\}$ into disjoint pairs $(f_i, f_{i'})$ for which $f_i(x) = f_{i'}(x)$ except at $x = v_r$. For such pairs,

$$\mathbb{1}_{[h(v_r) \neq f_i(v_r)]} = \mathbb{1}_{[h(v_r) \neq f_{i'}(v_r)]} = 1$$

such that the previous inequality becomes

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{P}_i}(A(S_j^i)) \geq \frac{1}{2}.$$

This, combined with the earlier results, implies

$$\max_{i \in [T]} \mathop{\mathbb{E}}_{S \sim \mathcal{P}_i^m} [L_{\mathcal{P}_i}(A(S))] \geq \frac{1}{4}. \qquad (\star)$$

Lemma B.1 in UML tells us that for any radom variable Z

$$\mathbb{P}[Z > a] = \frac{\mathbb{E}[Z] - a}{1 - a}.$$

Given our result ($\star$), we can apply this lemma to determine that for any learning algoithm A, there exists some distribution $\mathcal{P}$ such that

$$\mathbb{P}[L_{\mathcal{P}}(A(S)) > 1/8] = \frac{\mathbb{E}[L_{\mathcal{P}}(A(S))] - 1/4}{1 - 1/4} \geq \frac{1/4 - 1/8}{7/8} = 1/7.$$

2. Take $(d+1)$ vectors $x_k \in \mathbb{R}^d$, augment them each with a 1, and form the matrix $\tilde{X}$ for which

$$\tilde{X}^T \tilde{w} = \begin{bmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_{d+1}^T & 1 \end{bmatrix} \begin{bmatrix} \theta \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_{d+1} \end{bmatrix}$$

We can always pick each x_k such that $\tilde{X}$ is full rank in which case can we assign any values to y_k by setting θ, b properly. This means $\mathcal{H}$ shatters our set of $(d+1)$ vectors, since for any labeling h, we can always pick form a vector y for which $h(x_k) = y_k$ for all k.

Now if we form $\tilde{X}$ out of $(d+2)$ vectors, there will always be some x_j that is linearly dependent on the other vectors such that

$$x_j = \sum_{k \neq j} \alpha_k x_k.$$

But this means we have no control over the value of y_j. For any θ, b that we pick,

$$y_j = \tilde{x}_j \tilde{\theta} = \sum_{k \neq j} \alpha_k y_k$$

Then if we try to assign labels $\text{sign}(\alpha_k)$ to each x_k by setting each y_k accordingly we see that

$$y_j = \sum_{k \neq j} |\alpha_k y_k| \geq 0.$$

In such a case, $h(x_j)$ must equal 1 and so any size $(d+2)$ set cannot be shattered.

3. $\mathcal{H}_B$ for $d = 2$ represents the set of all axis-aligned halfspaces in $\mathbb{R}^2$. The result from question 2 tells us that, since $\mathcal{H}_B$ is a subset of the class of all halfspaces in $\mathbb{R}^2$, the VC dimension of $\mathcal{H}_B$ is at most 3. Furthermore, any triangle of 3 points in $\mathbb{R}^2$ can be shattered by $\mathcal{H}_B$ so the VC dimension is in fact 3.

Regarding the VC dimension of $\mathcal{H}$, we can obtain a lower bound as follows. Let g_r be a piece-wise constant function with at most $r+1$ pieces defined as

$$g_r(x) = \sum_{t=1}^{r+1} \alpha_t \mathbb{1}_{x \in (\theta_{t-1}, \theta_t]} \quad \alpha_i \in \{-1, 1\}$$

where $-\infty = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_r \leq \theta_{r+1} = \infty$. Let $\mathcal{G}_r$ represent the set of all such functions of a particular element x_i of the vector x and note that $\mathcal{G}_T \subset \mathcal{H}$. This can be seen by re-expressing g_T as

$$g_T(x_i) = \text{sign}\big(\sum_{t=1}^{T} w_t \ \text{sign}(x_i - \theta_t)\big) = \text{sign}\big(\sum_{t=1}^{T} w_t h(x_i, \theta_t, b_t)\big)$$

where $w_1 = 0.5$, $w_t = (-1)^t$ for $t > 1$, and $b_t = 1$ for all t. Then any set of $T+1$ points with unique positions along a single axis x_i can be labeled by some $g_t \in \mathcal{G}_T$ and is thus shattered by $\mathcal{H} \supset \mathcal{G}_T$. So the VC dimension of $\mathcal{H}$ is lower-bounded by $T+1$.