Exercise 1

$$S = \{(x_i, f(x_i))\}_{i=1}^m \subseteq (\mathbb{R}^d \times \{0,1\})^m$$

Consider polynomial $p_s(x) = -(x-x_1)^2(x-x_2)^2 \dots (x-x_m)^2$

Our $h_s(x)$ function is defined as

$$h_s(x) = \begin{cases} 1 & \text{if } \exists i \in \{1,2,\dots m\}, \text{ s.t. } x_i = x \\ 0 & \text{otherwise} \end{cases}$$

Let's show that $h_s(x) = 1$ if and only if $p_s(x) \geq 0$

$\boxed{i}$ first, let's show that $h_s(x) = 1 \Rightarrow p_s(x) \geq 0$

Consider x, such that $h_s(x) = 1$, then it means that $\exists\, i \in [m]$, s.t. $x_i = x$. Then, $(x - x_i)^2 = 0$ and, therefore, $p_s(x) = 0$. Hence, $p_s(x) \geq 0$.

$\boxed{ii}$ Now, let's show that $p_s(x) \geq 0 \Rightarrow h_s(x) = 1$

Consider x, such that $p_s(x) \geq 0$.

We know that for each $i \in [m]$, $(x - x_i)^2 \geq 0$ and equal to 0, if and only if $x = x_i$.

Assume in this case $x \neq x_i$ for all $i \in [m]$. Then, each of $(x - x_i)^2$ is greater than 0, hence $\prod_{i=1}^m (x-x_i)^2 > 0$.

Then, $p_s(x) = -(x-x_1)^2(x-x_2)^2 \dots (x-x_m)^2 < 0$, ~~which~~ and that leads to contradiction.

Hence, our assumption was wrong and $\exists\, i \in [m]$, s.t. $x_i = x$

Therefore, $h_s(x) = 1$.

We have shown that $h_s(x) = 1 \Rightarrow p_s(x) \geq 0$ and $p_s(x) \geq 0 \Rightarrow h_s(x) = 1$

Hence, $h_s(x) = 1$, if and only if $p_s(x) \geq 0$.

q. e. d.

Exercise 2

$$\underset{S|_x \sim D^m}{E}\left(L_S(h)\right) = \underset{S|_x \sim D^m}{E}\left(\frac{|\{i \in [m] : h(x_i) \neq f(x_i)\}|}{m}\right) =$$

$$= \frac{1}{m} \underset{S|_x \sim D^m}{E}\left(|\{i \in [m] : h(x_i) \neq f(x_i)\}|\right) =$$

$$= \frac{1}{m} \sum_{i=1}^{m} \underset{S|_x \sim D^m}{P}\left(h(x_i) \neq f(x_i)\right).$$

Since we have binary classifier, $\underset{S_x \sim D^m}{P}\left(h(x_i) \neq f(x_i)\right)$ is the same for all i. Then:

$$= \frac{1}{m} \sum_{i=1}^{m} \underset{x \sim D^m}{P}\left(h(x_{\xi}) \neq f(x_{\xi})\right) =$$

$$= \frac{1}{m} \cdot m \cdot \underset{x \sim D^m}{P}\left(h(x_{\xi}) \neq f(x_{\xi})\right) =$$

$$= \underset{x \sim D^m}{P}\left(h(x) \neq f(x)\right) =$$

$$= L_{D,f}(h).$$

q.e.d.

Exercise 3

1 Consider classifier h_1 obtained from algorithm A for sample S.

Also, consider classifier $h^* \in \mathcal{H}^2_{rec}$, s.t. $L_{(D,f)}(h^*) = 0$.
It exists due to the Realizability Assumption.

Then, $L_S(h^*)$ is also equal to 0. Therefore, all positive examples are contained within rectangle of classifier h^*. Since rectangle of h_1 is the smallest rectangle enclosing all positive examples, it must be located within borders of h^* rectangle.

Since $L_S(h^*) = 0$, all negative examples are outside of the h^* rectangle and, therefore, they are outside of the h_1 rectangle.

Since all negative examples are outside of the h_1 rectangle and all positive are inside, we get that ~~h~~ $L_S(h_1) = 0$, hence A is an ERM.

$\boxed{\text{Exercise 3}}\boxed{\text{cont.}}$

$\boxed{2}$ Since R^* is the rectangle that generates the labels, all positive examples are inside R^* and all negative are outside.

Assume $R(S)$ is not fully contained within R^*. Then, consider $R(S) \cap R^*(S)$. Since all positives are contained within both $R(S)$ and $R^*(S)$ and all negatives are outside of both $R(S)$ and $R^*(S)$, $R(S) \cap R^*(S)$ is a rectangle enclosing all positive examples in the training set. But we know that $R(S)$ is the smallest such rectangle, since it is obtained by A. Hence, $R(S) \cap R^*(S)$ and $R(S)$ are of the same size, and that means that $R(S) \subseteq R^*(S)$.

Now, if S has positive examples in all of the R_1, R_2, R_3, R_4, let's show that A has error at most ε.

Let $R(S) = R(a_1', b_1', a_2', b_2')$. Since it has <u>all</u> positive examples and each of R_1, R_2, R_3, R_4 have at least one, we get that $a_1 \geq a_1' \geq a_1^*$, $a_2 \geq a_2' \geq a_2^*$, $b_1 \leq b_1' \leq b_1^*$ and $b_2 \leq b_2' \leq b_2^*$. Since probability mass of $R(a_1, b_1, a_2, b_2)$ is at least probability mass of R^* minus sum of probability masses of R_1, R_2, R_3, R_4, we get that probability mass of the rectangle $R(a_1, b_1, a_2, b_2)$ is at least $1-\varepsilon$. Since $R(a_1, b_1, a_2, b_2)$ is fully contained within $R(a_1', b_1', a_2', b_2')$, probab. mass of $R(S)$ is at least $1-\varepsilon$, hence hypothesis returned by A has error of at most ε.

Let m be the size of the training set. Then for each point in the set, the probability that it is not contained within R_i is $1-\frac{\varepsilon}{4}$ for each of $i \in \{1,2,3,4\}$. Hence, for each $i \in \{1,\dots 4\}$, the probability that S doesn't contain an example in R_i will be $(1-\frac{\varepsilon}{4})^m$.

Then, probability that S doesn't contain ~~exa~~ positive example in at least one of R_i, $i \in \{1,..4\}$ will be equal to $4(1-\frac{\varepsilon}{4})^m$, which is smaller ~~than~~ or equal to $4 \cdot e^{(-\frac{\varepsilon}{4}) \cdot m}$.

Hence, the probability that the hypothesis returned by A has error of at most ε is larger ~~smaller~~ or equal to $1-4e^{(-\frac{\varepsilon}{4})m}$. Let's denote this probability ~~as~~ $1-\sigma$.

Then, $\sigma \leq 4e^{-\frac{\varepsilon}{4} \cdot m}$ gives us inequality

$$m \geq \frac{4\log(\frac{4}{\sigma})}{\varepsilon}.$$

Hence, if A receives a training set of size $m \geq \frac{4\log(4/\sigma)}{\varepsilon}$, then with probability of at least $1-\sigma$ it returns a hypothesis with error of at most ε.

$\boxed{3}$ Let $R^* = R^d(a_1^*, b_1^*, a_2^*, b_2^*, \dots a_d^*, b_d^*)$ be d-dimension rectangle that generates the labels and let f be the corresponding hypothesis.

Let $a_1 \geq a_1^*$ be a number, s.t. prob. mass of the rectangle $R_1 = R^d(a_1^*, a_1, a_2^*, b_2^*, \dots a_d^*, b_d^*)$ exactly $\frac{\varepsilon}{2d}$. Similarly, let $a_2, a_3 \dots, a_d$ be numbers such that prob. masses of $R_2 = R^d(a_1^*, b_1^*, a_2^*, a_2, \dots a_d^*, b_d^*)$, ..., $R_d = R^d(a_1^*, b_1^*, \dots a_d^*, a_d)$ are exactly $\frac{\varepsilon}{2d}$. And let $b_1, b_2, \dots, b_d$ be such numbers that prob. masses of $R_{d+1} = R^d(b_1, b_1^*, a_2^*, b_2^*, \dots a_d^*, b_d^*)$, ... $R_{2d} = (a_1^*, b_1^*, \dots a_d^*, b_d^*)$ are all exactly $\frac{\varepsilon}{2d}$. Let $R(S)$ be the rectangle generated by A.

Similarly to $\mathbb{R}^2$ case, we can show that $R(S) \subseteq R^*$. Since probab. masses of each of R_i, $i \in [2d]$ is exactly $\frac{\varepsilon}{2d}$, probability mass of the rectangle $R^d(a_1, b_1, a_2, b_2, \dots a_d, b_d)$ is at least $1-\varepsilon$. ~~Since~~

If S contains positive examples in all of the R_i, then $R^d(a_1, b_1, \dots a_d, b_d)$ will be fully contained within $R(S)$ and, therefore, $R(S)$ will have the probability mass of at least $1-\varepsilon$.

Hence, hypothesis returned by will have an error of at most ε.

Now, let m be the size of the training set. Then, for each point in the set, probability that it is not contained within R_i is $1-\frac{\varepsilon}{2d}$ for each $i \in [2d]$. Hence, for each $i \in [2d]$, the probab. that S doesn't contain an ~~exam~~ example in R_i will be $(1-\frac{\varepsilon}{2d})^m$. Then, probab. that S doesn't contain positive example in at least one of R_i, $i \in [2d]$, will be equal to $2d(1-\frac{\varepsilon}{2d})^m$, which is smaller or equal to $2d\, e^{(-\frac{\varepsilon}{2d})m}$.

Hence, the probab. that the hypothesis returned by A has error of at most ε is larger or equal to $1-2d\, e^{(-\frac{\varepsilon}{2d})m}$. Let's denote this probability $1-\sigma$.

Then, $\sigma \leq 4e^{-\frac{\varepsilon}{2d}m}$ and, therefore, $m \geq \frac{2d \log(\frac{2d}{\sigma})}{\varepsilon}$

Hence, if A receives a training set of the size $m \geq \frac{2d\log(2d/\sigma)}{\varepsilon}$, then with probability of at least $1-\sigma$, it returns a hypothesis with error of at most ε.