Exercise 1

(1) First, let's show that given $\delta \in (0,1)$ and given $0 < \varepsilon_1 \leq \varepsilon_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta)$

Since $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$, we get that given a sample of size $m_1 \geq m_{\mathcal{H}}(\varepsilon_1, \delta)$, we have that with probability of at least $1-\delta$ over $S \sim D^{m_1}$, $L_{D,f}(A(S)) < \varepsilon_1$.

Since $\varepsilon_1 \leq \varepsilon_2$, for the sample of size $m_1 \geq m_{\mathcal{H}}(\varepsilon_1, \delta)$ we also have that w.p. of at least $1-\delta$ over $S \sim D^{m_1}$, $L_{D,f}(A(S)) < \varepsilon_2$. Hence, the sample size m_2 required for $L_{D,f}(A(S))$ to be smaller than ε_2 w.p. at least $1-\delta$ over $S \sim D^{m_2}$ is guaranteed to work, if $m_2 \geq m_{\mathcal{H}}(\varepsilon_1, \delta)$. Therefore, the sample complexity for ε_2 is at most $m_{\mathcal{H}}(\varepsilon_1, \delta)$. Hence, we have shown that $m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta)$

(2) Now, let's show that given $\varepsilon \in (0,1)$ and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\varepsilon, \delta_1) \geq m_{\mathcal{H}}(\varepsilon, \delta_2)$

Since $\mathcal{H}$ is PAC-learnable, given any $m \geq m(\varepsilon, \delta_1)$, we have $P_{S\sim D^m}(A(S) < \varepsilon) \geq 1-\delta_1$. Since $\delta_1 \leq \delta_2$, we have that for any $m \geq m(\varepsilon, \delta_1)$, $P_{S\sim D^m}(A(S) < \varepsilon) \geq 1-\delta_1 \geq 1-\delta_2$

Therefore, the sample complexity for δ_2 is at most $m_{\mathcal{H}}(\varepsilon, \delta_1)$

Hence, $m_{\mathcal{H}}(\varepsilon, \delta_1) \geq m(\varepsilon, \delta_2)$

By (1) and (2) we get that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters.

q.e.d.

Exercise 2

1. Algorithm:

Go through every x in the sample and check whether it belongs to our discrete domain $\mathcal{X}$.

If we find such x, then function h_z, where $z = x$ is the outcome of our algorithm. Due to realizability assumption, there can only be one such x in the sample and, therefore, $L_S(h_z)$ will be 0.

If we don't find such x, then h^- is the outcome of our algorithm. Since in this case we don't have any positives in the sample, $L_S(h^-)$ will be 0.

2. If the algorithm described above outputs h_z, then, due to the realizability assumption it will mean that our hypothesis h_z correctly identifies the only positive in our domain, and hence, $L_{D,f}(h_z)$ will be 0, which is less than ε for all $\varepsilon > 0$.

The only case in which our algorithm won't identify the correct labeling function is when the only positive in distribution (let's call it x_j) is not selected to our sample S_x and our algorithm incorrectly returns h^-. Since all items in the sample are i.i.d. selected, the probability of this happening is $(1 - P(x_j))^m$, where m is the size of our sample. We want this probability to be at most δ $(\delta > 0)$, in order for $\mathcal{H}_{singleton}$ to be PAC-learnable

Since in this case, the population risk will be equal to $P_{\mathcal{D}}(x_j')$, we get the inequality for the size of the sample m:

$$(1-\varepsilon)^m \leq \delta$$

$$\Updownarrow$$

$$m \geq \lceil \log_{(1-\varepsilon)} \delta \rceil$$

the required upper bound on the sample complexity

Exercise 3 . Fix some distribution D over X.

Assuming realizability, ~~Assumin~~ Let R^* with radius r^* be the concentric circle that generates the labels and let f be the corresponding hypothesis. Let $r < r^*$ be ~~such that~~ one radius of the concentric circle R, such that the probability mass of the area between R and R^* is exactly ε. (Let's denote that area R')

~~If R' has a positive example from the samples S~~

Let A be the algorithm that returns the smallest ~~rectangle~~ circle enclosing all positive examples in the training set. From realizability assumption we get that A does not mislable any negative examples (if it does, there should be smaller circle enclosing all positive examples, which contradicts definition of $A(S)$). Hence, A is an ERM.

Since $A(S)$ is the smallest circle enclosing all positive examples, $A(S) \subseteq R^*$. If R' contains any positive example, then the boundary of A lies between R^* and R. Since the probability mass of that region is ε, empirical risk of $A(S)$ in this case will be at most ε.

The probability of positive example being inside R is $1-\varepsilon$, hence the probability of all positive samples ~~to~~ be inside R is $(1-\varepsilon)^m$, where m is the size of the sample. Therefore, the probability that at least one positive is in R' is $1-(1-\varepsilon)^m$.

We get that $P(L_{D,f}(A(S)) < \varepsilon) \geq 1-(1-\varepsilon)^m \geq 1-e^{-m\varepsilon}$

δ

Rewriting this inequality, we get $m \geq \frac{\log(1/\delta)}{\varepsilon}$

Therefore, we have show that $\exists$ algorithm A, s.t. $\forall\ \varepsilon, \delta > 0$, $\forall$ distributions D, $\forall$ labeling fcns f, s.t. D, f realizable by $\mathcal{H}$, given $m \geq \lceil \frac{\log(1/\delta)}{\varepsilon} \rceil$, then w.p. at least $1-\delta$ over $S \sim D^m$, we have

$$L_{D,f}(A(S)) < \varepsilon$$

Hence, $\mathcal{H}$ is PAC-learnable and its sample complexity is bounded by $m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil \frac{\log(1/\delta)}{\varepsilon} \rceil$

q. e. d.

Exercise 6

Let $\mathcal{H}$ be agnostic PAC-learnable and let A be a successful agnostic PAC learner for $\mathcal{H}$.

Then, ~~$\exists$~~ consider $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$, s.t. $\forall \varepsilon, \delta > 0$, $\forall D$ on $X \times Y$, given $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ w.p. $\geq (1-\delta)$ over $S \sim D^m$ it is true that $L_D(A(S)) \leq \inf_{h \in \mathcal{H}} L_D(h) + \varepsilon$

Using definition (3.1 from the book) of the true error, we get that $L_D(h) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$

Therefore, given $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$, we know that:

$$\star 1 \quad P_{S\sim D^m}\left[L_D(A(S)) \leq \inf_{h \in \mathcal{H}} P_{(x,y)\sim D}[h(x) \neq y] + \varepsilon\right] \geq (1-\delta)$$

For the binary classifier loss function is defined as following:

$$\ell(h,(x,y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$$

Hence, $\mathop{E}_{(x,y)\sim D}[\ell(h,(x,y))] = 0 \cdot \mathop{P}_{(x,y)\sim D}[h(x)=y] + 1 \cdot \mathop{P}_{(x,y)\sim D}[h(x) \neq y] =$

$$= \mathop{P}_{(x,y)\sim D}[h(x) \neq y].$$

Therefore, for our algorithm A and function $m_{\mathcal{H}}(\cdot,\cdot)$, we get that:

$$\star 2 \quad \mathop{P}_{S\sim D^m}\left[L_D(A(S)) \leq \inf_{h\in\mathcal{H}} \mathop{E}_{(x,y)\sim D}[\ell(h,(x,y))] + \varepsilon\right] \geq (1-\delta)$$

Hence, $\mathcal{H}$ is PAC-learnable and A is a successful PAC learner for $\mathcal{H}$ by definition.

Exercise 7

By definition 3.1 of the true risk, we get that the true risk for randomly chosen classifier g is equal to:

$$L_D(g) = \underset{(x,y)\sim D}{P}[g(x) \neq y] = \begin{cases} P(y=1|x), & \text{if } g(x)=0 \\ P(y=0|x), & \text{if } g(x)=1 \end{cases}$$

Predictor is optimal, if it has the lowest possible value of $L_D(g)$.

As we can see from the expression above, $L_D(g)$ has the lowest possible value, when the predictor is such that $L_D(g) = P(y=1|x)$, when $P(y=1|x) \leq P(y=0|x)$ and $L_D(g) = P(y=0|x)$ when $P(y=0|x) \leq P(y=1|x)$.

Now, consider $f_D(x)$ (Bayes Optimal Predictor)

$$L_D(f_D(x)) = \begin{cases} P(y=1|x), & \text{if } f_D(x)=0 \\ P(y=0|x), & \text{if } f_D(x)=1 \end{cases} = \text{(by definition)}$$

$$= \begin{cases} P(y=1|x), & \text{if } P(y=1|x) < \frac{1}{2} \\ P(y=0|x), & \text{if } P(y=1|x) \geq \frac{1}{2} \end{cases} = \begin{cases} P(y=1|x), & \text{if } P(y=1|x) < \frac{1}{2} \\ P(y=0|x), & \text{if } P(y=0|x) \leq \frac{1}{2} \end{cases} =$$

$$= \begin{cases} P(y=1|x), & \text{if } P(y=1|x) < P(y=0|x) \\ P(y=0|x), & \text{if } P(y=0|x) < P(y=1|x) \end{cases}$$

(since $P(y=1|x) + P(y=0|x) = 1$)

As we can see, $L_D(g)$ has the lowest possible value when $g = f_D$

Hence, $\forall$ classifier g from X to $\{0,1\}$, $L_D(f_D) \leq L_D(g)$

q.e.d.