**Exercise 3.1**

The hypothesis class $\mathcal{H}$ being PAC learnable with sample complexity $m(\epsilon, \delta)$ means that there is a learning algorithm $A$ such that when running $A$ on $m \geq m(\epsilon, \delta)$ i.i.d. samples generated by $D$ and labeled by $f$, with probability at least $1 - \delta$, $A$ returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon$.

Given $0 < \epsilon_1 \leq \epsilon_2 < 1$, consider $m \geq m(\epsilon_1, \delta)$, we have that with probability at least $1 - \delta$, $A$ returns a hypothesis $h \in \mathcal{H}$ with $L_{D,f}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that $m(\epsilon_1, \delta)$ is a sufficient number of samples for accuracy $\epsilon_2$. Therefore, $m(\epsilon_1, \delta) \geq m(\epsilon_2, \delta)$.

The proof of $m(\epsilon, \delta_1) \geq m(\epsilon, \delta_2)$ for $0 < \delta_1 \leq \delta_2 < 1$ follows analogously from the definition.

**Exercise 3.3**

The realizability assumption for $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$ implies that there is a circle such that any $x$ inside it has label $y = 1$, and the learning task here is to distinguish this circle. Now consider an ERM algorithm which given a training sequence $S = \{(x_i, y_i)\}_{i=1}^m$, returns the hypothesis $h$ corresponding to the tightest circle which contains all the positive instances in $S$ where $y_i = 1$ and does not allow false negative predictions. With the realizability assumption let $h^*$ be the circle with zero training error and $r^*$ be the corresponding radius.

Let $\tilde{r} \leq r^*$ be a scalar such that $\mathbb{P}_{x \sim D}(x : \tilde{r} \leq \|x\| \leq r^*) = \epsilon$ and $E = \{x \in \mathbb{R}^2 : \tilde{r} \leq \|x\| \leq r^*\}$. We have

$$
\mathbb{P}(L_{D}(h_S) \geq \epsilon) \leq \mathbb{P}(\text{no points in } S \text{ belongs to } E)
= (1 - \epsilon)^m
\leq e^{-\epsilon m}
$$

The desired bound on the sample complexity follows from requiring $e^{-\epsilon m} \leq \delta$.

**Exercise 3.7**

Let $g$ be any (potentially probabilistic) classifier from $\mathcal{X}$ to $\{0, 1\}$. Note that for 0-1 loss

$$
L_{D}(g) = \mathbb{E}_{(x,y) \sim D}[\mathbf{1}_{g(x) \neq y}] = \mathbb{E}_{x \sim D}[\mathbb{E}_{y \sim D_{Y|x}}[\mathbf{1}_{g(x) \neq y}]] = \mathbb{E}_{x \sim D}[\mathbb{P}_{y \sim D_{Y|x}}(g(X) \neq Y | X = x)],
$$

$$
L_{D}(f_D) = \mathbb{E}_{x \sim D}[\mathbb{P}_{y \sim D_{Y|x}}(f_D(X) \neq Y | X = x)].
$$
We should compare the two conditional probabilities inside the expectation. Let \( x \in \mathcal{X} \) and \( a_x = \mathbb{P}(Y = 1|X = x) \). We have

\[
\mathbb{P}(g(X) \neq Y|X = x) = \mathbb{P}(g(X) = 0|X = x) \cdot \mathbb{P}(Y = 1|X = x) \\
+ \mathbb{P}(g(X) = 1|X = x) \cdot \mathbb{P}(Y = 0|X = x) \\
= \mathbb{P}(g(X) = 0|X = x) \cdot a_x + \mathbb{P}(g(X) = 1|X = x) \cdot (1 - a_x) \\
\geq \mathbb{P}(g(X) = 0|X = x) \cdot \min\{a_x, 1 - a_x\} \\
+ \mathbb{P}(g(X) = 1|X = x) \cdot \min\{a_x, 1 - a_x\} \\
= \min\{a_x, 1 - a_x\}.
\]

When \( g = f_D \) we should replace \( \mathbb{P}(g(X) = 0|X = x) \) by \( 1_{a_x < 1/2} \) and \( \mathbb{P}(g(X) = 1|X = x) \) by \( 1_{a_x \geq 1/2} \). Then the above inequality is tight:

\[
\mathbb{P}(f_D(X) \neq Y|X = x) = 1_{a_x < 1/2} \cdot a_x + 1_{a_x \geq 1/2} \cdot (1 - a_x) = \min\{a_x, 1 - a_x\}.
\]

Therefore, we have \( L_D(f_D) \leq L_D(g) \).

**Exercise 3.8**

1. Solved already in Exercise 3.7.

2. We have shown in Exercise 3.7 that the Bayes optimal predictor \( f_D \) is optimal w.r.t. \( \mathcal{D} \); in other words, \( f_D \) is always better than any other learning algorithm w.r.t. \( \mathcal{D} \).

3. Take \( \mathcal{D} \) to be any probability distribution and \( B = f_D \).

**Exercise 4.1**

1 \( \Rightarrow \) 2: Assume for every \( \epsilon, \delta > 0 \) there exists \( m(\epsilon, \delta) \) such that \( \forall m \geq m(\epsilon, \delta) \)

\[
\mathbb{P}_{S \sim D^m}(L_D(A(S)) > \epsilon) < \delta. \quad (1)
\]

Then using the definition of expectation

\[
\mathbb{E}_{S \sim D^m}[L_D(A(S))] \leq \mathbb{P}_{S \sim D^m}(L_D(A(S)) > \epsilon) \cdot 1 + \mathbb{P}_{S \sim D^m}(L_D(A(S)) \leq \epsilon) \cdot \epsilon \\
\leq \mathbb{P}_{S \sim D^m}(L_D(A(S)) > \epsilon) + \epsilon \\
\leq \delta + \epsilon,
\]

where the last inequality follows from the assumption (1). Now set \( \delta = \epsilon \). We have for every \( \epsilon > 0 \) there exists \( m(\epsilon, \epsilon) \) such that \( \forall m \geq m(\epsilon, \epsilon) \)

\[
\mathbb{E}_{S \sim D^m}[L_D(A(S))] \leq 2\epsilon. \quad (2)
\]

So it is valid to pass both sides of (2) to the limit \( \lim_{m \to \infty} \lim_{\epsilon \to 0} \), which gives

\[
\lim_{m \to \infty} \mathbb{E}_{S \sim D^m}[L_D(A(S))] \leq 0.
\]

Also by definition \( \mathbb{E}_{S \sim D^m}[L_D(A(S))] \geq 0 \). Thus we conclude \( \lim_{m \to \infty} \mathbb{E}_{S \sim D^m}[L_D(A(S))] = 0 \).
2 ⇒ 1: Assume that \( \lim_{m \to \infty} \mathbb{E}_{S \sim D^m}[L_D(A(S))] = 0 \). For every \( \epsilon, \delta \in (0, 1) \) there exists some \( m_0 \in \mathbb{N} \) such that for every \( m \geq m_0 \), \( \mathbb{E}_{S \sim D^m}[L_D(A(S))] \leq \epsilon \delta \). By Markov’s inequality,

\[
\mathbb{P}_{S \sim D^m}(L_D(A(S)) > \epsilon) \leq \frac{\mathbb{E}_{S \sim D^m}[L_D(A(S))]}{\epsilon} \leq \frac{\epsilon \delta}{\epsilon} = \delta.
\]

**Exercise 4.2**

Using Hoeffding’s inequality on \( L_D \in [a, b] \) we have

\[
\mathbb{P}_{S \sim D^m}(|L_D(h) - L_S(h)| > \epsilon) \leq 2 \exp \left( -\frac{2m\epsilon^2}{(b-a)^2} \right).
\]

Then we substitute this into the step where the union bound is used:

\[
\mathbb{P}_{S \sim D^m}(\exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon) \leq \sum_{h \in \mathcal{H}} \mathbb{P}_{S \sim D^m}(|L_D(h) - L_S(h)| > \epsilon) \leq 2|\mathcal{H}| \exp \left( -\frac{2m\epsilon^2}{(b-a)^2} \right).
\]

The desired bound on the sample complexity follows from requiring \( 2|\mathcal{H}| \exp \left( -\frac{2m\epsilon^2}{(b-a)^2} \right) \leq \delta \).