2 Gentle Start

1. Given $S = ((x_i, y_i))_{i=1}^m$, define the multivariate polynomial

$$p_S(x) = - \prod_{i \in [m]: y_i = 1} \|x - x_i\|^2.$$ 

Then, for every $i$ s.t. $y_i = 1$ we have $p_S(x_i) = 0$, while for every other $x$ we have $p_S(x) < 0$.

2. By the linearity of expectation,

$$\mathbb{E}_{S | x \sim D^m}[L_S(h)] = \mathbb{E}_{S | x \sim D^m} \left[ \frac{1}{m} \sum_{i=1}^m \mathbb{I}_{[h(x_i) \neq f(x_i)]} \right] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{x_i \sim D} [\mathbb{I}_{[h(x_i) \neq f(x_i)]}] = \frac{1}{m} \sum_{i=1}^m \mathbb{P}_{x_i \sim D} [h(x_i) \neq f(x_i)] = \frac{1}{m} \cdot m \cdot L_{(\mathcal{D}, f)}(h) = L_{(\mathcal{D}, f)}(h).$$

*The solutions to Chapters 13,14 were written by Shai Shalev-Shwartz*
3. (a) First, observe that by definition, \( A \) labels positively all the positive instances in the training set. Second, as we assume realizability, and since the tightest rectangle enclosing all positive examples is returned, all the negative instances are labeled correctly by \( A \) as well. We conclude that \( A \) is an ERM.

(b) Fix some distribution \( \mathcal{D} \) over \( \mathcal{X} \), and define \( R^* \) as in the hint. Let \( f \) be the hypothesis associated with \( R^* \) a training set \( S \), denote by \( R(S) \) the rectangle returned by the proposed algorithm and by \( A(S) \) the corresponding hypothesis. The definition of the algorithm \( A \) implies that \( R(S) \subseteq R^* \) for every \( S \). Thus,

\[
L_{(\mathcal{D},f)}(R(S)) = \mathcal{D}(R^* \setminus R(S))
\]

Fix some \( \epsilon \in (0,1) \). Define \( R_1, R_2, R_3 \) and \( R_4 \) as in the hint. For each \( i \in [4] \), define the event

\[
F_i = \{ S | x : S|_x \cap R_i = \emptyset \}
\]

Applying the union bound, we obtain

\[
\mathcal{D}^m(\{ S : L_{(\mathcal{D},f)}(A(S)) > \epsilon \}) \leq \mathcal{D}^m \left( \bigcup_{i=1}^{4} F_i \right) \leq \sum_{i=1}^{4} \mathcal{D}^m(F_i)
\]

Thus, it suffices to ensure that \( \mathcal{D}^m(F_i) \leq \delta/4 \) for every \( i \). Fix some \( i \in [4] \). Then, the probability that a sample is in \( F_i \) is the probability that all of the instances don’t fall in \( R_i \), which is exactly \((1 - \epsilon/4)^m\). Therefore,

\[
\mathcal{D}^m(F_i) = (1 - \epsilon/4)^m \leq \exp(-m\epsilon/4)
\]

and hence,

\[
\mathcal{D}^m(\{ S : L_{(\mathcal{D},f)}(A(S)) > \epsilon \}) \leq 4 \exp(-m\epsilon/4)
\]

Plugging in the assumption on \( m \), we conclude our proof.

(c) The hypothesis class of axis aligned rectangles in \( \mathbb{R}^d \) is defined as follows. Given real numbers \( a_1 \leq b_1, a_2 \leq b_2, \ldots, a_d \leq b_d \), define the classifier \( h_{(a_1,b_1,\ldots,a_d,b_d)} \) by

\[
h_{(a_1,b_1,\ldots,a_d,b_d)}(x_1, \ldots, x_d) = \begin{cases} 
1 & \text{if } \forall i \in [d], \ a_i \leq x_i \leq b_i \\
0 & \text{otherwise}
\end{cases}
\]
The class of all axis-aligned rectangles in $\mathbb{R}^d$ is defined as

$$\mathcal{H}_{\text{rec}}^d = \{ h_{(a_1,b_1,\ldots,a_d,b_d)} : \forall i \in [d], a_i \leq b_i, \}.$$ 

It can be seen that the same algorithm proposed above is an ERM for this case as well. The sample complexity is analyzed similarly. The only difference is that instead of 4 strips, we have $2d$ strips (2 strips for each dimension). Thus, it suffices to draw a training set of size $\lceil \frac{2d \log(2d/\delta)}{\epsilon} \rceil$.

(d) For each dimension, the algorithm has to find the minimal and the maximal values among the positive instances in the training sequence. Therefore, its runtime is $O(md)$. Since we have shown that the required value of $m$ is at most $\lceil \frac{2d \log(2d/\delta)}{\epsilon} \rceil$, it follows that the runtime of the algorithm is indeed polynomial in $d, 1/\epsilon$, and $\log(1/\delta)$.

3 A Formal Learning Model

1. The proofs follow (almost) immediately from the definition. We will show that the sample complexity is monotonically decreasing in the accuracy parameter $\epsilon$. The proof that the sample complexity is monotonically decreasing in the confidence parameter $\delta$ is analogous.

Denote by $D$ an unknown distribution over $\mathcal{X}$, and let $f \in \mathcal{H}$ be the target hypothesis. Denote by $A$ an algorithm which learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot, \cdot)$. Fix some $\delta \in (0, 1)$. Suppose that $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. We need to show that $m_1 \overset{\text{def}}{=} m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta) \overset{\text{def}}{=} m_2$. Given an i.i.d. training sequence of size $m \geq m_1$, we have that with probability at least $1 - \delta$, $A$ returns a hypothesis $h$ such that

$$L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2.$$ 

By the minimality of $m_2$, we conclude that $m_2 \leq m_1$.

2. (a) We propose the following algorithm. If a positive instance $x_+$ appears in $S$, return the (true) hypothesis $h_{x_+}$. If $S$ doesn’t contain any positive instance, the algorithm returns the all-negative hypothesis. It is clear that this algorithm is an ERM.

(b) Let $\epsilon \in (0,1)$, and fix the distribution $\mathcal{D}$ over $\mathcal{X}$. If the true hypothesis is $h^-$, then our algorithm returns a perfect hypothesis.
Assume now that there exists a unique positive instance \( x_+ \). It’s clear that if \( x_+ \) appears in the training sequence \( S \), our algorithm returns a perfect hypothesis. Furthermore, if \( D[\{x_+\}] \leq \epsilon \) then in any case, the returned hypothesis has a generalization error of at most \( \epsilon \) (with probability 1). Thus, it is only left to bound the probability of the case in which \( D[\{x_+\}] > \epsilon \), but \( x_+ \) doesn’t appear in \( S \). Denote this event by \( F \). Then

\[
P_{S|x\sim D^m}(F) \leq (1 - \epsilon)^m \leq e^{-m\epsilon}.
\]

Hence, \( \mathcal{H}_{\text{Singleton}} \) is PAC learnable, and its sample complexity is bounded by

\[
m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil.
\]

3. Consider the ERM algorithm \( A \) 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. Denote the radius of this hypothesis by \( \hat{r} \). Assume realizability and let \( h^\star \) be a circle with zero generalization error. Denote its radius by \( r^\star \).

Let \( \epsilon, \delta \in (0, 1) \). Let \( \bar{r} \leq r^\star \) be a scalar s.t. \( D_X(\{x : \bar{r} \leq \|x\| \leq r^\star\}) = \epsilon \). Define \( E = \{x \in \mathbb{R}^2 : \bar{r} \leq \|x\| \leq r^\star\} \). The probability (over drawing \( S \)) that \( L_D(h_S) \geq \epsilon \) is bounded above by the probability that no point in \( S \) belongs to \( E \). This probability of this event is bounded above by

\[
(1 - \epsilon)^m \leq e^{-cm}.
\]

The desired bound on the sample complexity follows by requiring that \( e^{-cm} \leq \delta \).

4. We first observe that \( \mathcal{H} \) is finite. Let us calculate its size accurately. Each hypothesis, besides the all-negative hypothesis, is determined by deciding for each variable \( x_i \), whether \( x_i, \bar{x}_i \) or none of which appear in the corresponding conjunction. Thus, \( |\mathcal{H}| = 3^d + 1 \). We conclude that \( \mathcal{H} \) is PAC learnable and its sample complexity can be bounded by

\[
m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{d\log 3 + \log(1/\delta)}{\epsilon} \right\rceil.
\]

Let’s describe our learning algorithm. We define \( h_0 = x_1 \cap \bar{x}_1 \cap \ldots \cap x_d \cap \bar{x}_d \). Observe that \( h_0 \) is the always-minus hypothesis. Let \( ((a^1, y^1), \ldots, (a^m, y^m)) \) be an i.i.d. training sequence of size \( m \). Since
we cannot produce any information from negative examples, our algorithm neglects them. For each positive example \( a \), we remove from \( h \) all the literals that are missing in \( a \). That is, if \( a_i = 1 \), we remove \( \bar{x}_i \) from \( h \) and if \( a_i = 0 \), we remove \( x_i \) from \( h \). Finally, our algorithm returns \( h_m \).

By construction and realizability, \( h_i \) labels positively all the positive examples among \( a_1, \ldots, a_i \). From the same reasons, the set of literals in \( h_i \) contains the set of literals in the target hypothesis. Thus, \( h_i \) classifies correctly the negative elements among \( a_1, \ldots, a_i \). This implies that \( h_m \) is an ERM.

Since the algorithm takes linear time (in terms of the dimension \( d \)) to process each example, the running time is bounded by \( O(m \cdot d) \).

5. Fix some \( h \in \mathcal{H} \) with \( L_{D_m,f}(h) > \epsilon \). By definition,

\[
\sum_{i=1}^{m} \mathbb{P}_{X \sim D_i} [h(X) = f(X)] < 1 - \epsilon.
\]

We now bound the probability that \( h \) is consistent with \( S \) (i.e., that \( L_S(h) = 0 \)) as follows:

\[
P_{S \sim \prod_{i=1}^{m} D_i} [L_S(h) = 0] = \prod_{i=1}^{m} \mathbb{P}_{X \sim D_i} [h(X) = f(X)]
\]

\[
= \left( \prod_{i=1}^{m} \mathbb{P}_{X \sim D_i} [h(X) = f(X)] \right)^{\frac{1}{m}}
\]

\[
\leq \left( \sum_{i=1}^{m} \mathbb{P}_{X \sim D_i} [h(X) = f(X)] \right)^{m}
\]

\[
< (1 - \epsilon)^m
\]

\[
\leq e^{-\epsilon m}.
\]

The first inequality is the geometric-arithmetic mean inequality. Applying the union bound, we conclude that the probability that there exists some \( h \in \mathcal{H} \) with \( L_{D_m,f}(h) > \epsilon \), which is consistent with \( S \) is at most \( |\mathcal{H}| \exp(-\epsilon m) \).

6. Suppose that \( \mathcal{H} \) is agnostic PAC learnable, and let \( A \) be a learning algorithm that learns \( \mathcal{H} \) with sample complexity \( m_{\mathcal{H}}(\cdot, \cdot) \). We show that \( \mathcal{H} \) is PAC learnable using \( A \).
Let $\mathcal{D}, f$ be an (unknown) distribution over $\mathcal{X}$, and the target function respectively. We may assume w.l.o.g. that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0, 1\}$, where the conditional probability of $y$ given $x$ is determined deterministically by $f$. Since we assume realizability, we have $\inf_{h \in H} L_{\mathcal{D}}(h) = 0$. Let $\epsilon, \delta \in (0, 1)$. Then, for every positive integer $m \geq m_H(\epsilon, \delta)$, if we equip $A$ with a training set $S$ consisting of $m$ i.i.d. instances which are labeled by $f$, then with probability at least $1 - \delta$ (over the choice of $S | x$), it returns a hypothesis $h$ with

$$L_{\mathcal{D}}(h) \leq \inf_{h' \in H} L_{\mathcal{D}}(h') + \epsilon$$

$$= 0 + \epsilon$$

$$= \epsilon .$$

7. Let $x \in \mathcal{X}$. Let $\alpha_x$ be the conditional probability of a positive label given $x$. We have

$$\mathbb{P}[f_{\mathcal{D}}(X) \neq y | X = x] = \mathbb{I}_{[\alpha_x \geq 1/2]} \cdot \mathbb{P}[Y = 0 | X = x] + \mathbb{I}_{[\alpha_x < 1/2]} \cdot \mathbb{P}[Y = 1 | X = x]$$

$$= \mathbb{I}_{[\alpha_x \geq 1/2]} \cdot (1 - \alpha_x) + \mathbb{I}_{[\alpha_x < 1/2]} \cdot \alpha_x$$

$$= \min\{\alpha_x, 1 - \alpha_x\} .$$

Let $g$ be a classifier\(^1\) from $\mathcal{X}$ to $\{0, 1\}$. We have

$$\mathbb{P}[g(X) \neq Y | X = x] = \mathbb{P}[g(X) = 0 | X = x] \cdot \mathbb{P}[Y = 1 | X = x] + \mathbb{P}[g(X) = 1 | X = x] \cdot \mathbb{P}[Y = 0 | X = x]$$

$$= \mathbb{P}[g(X) = 0 | X = x] \cdot \alpha_x + \mathbb{P}[g(X) = 1 | X = x] \cdot (1 - \alpha_x)$$

$$\geq \mathbb{P}[g(X) = 0 | X = x] \cdot \min\{\alpha_x, 1 - \alpha_x\}$$

$$+ \mathbb{P}[g(X) = 1 | x] \cdot \min\{\alpha_x, 1 - \alpha_x\}$$

$$= \min\{\alpha_x, 1 - \alpha_x\} ,$$

The statement follows now due to the fact that the above is true for every $x \in \mathcal{X}$. More formally, by the law of total expectation,

$$L_{\mathcal{D}}(f_{\mathcal{D}}) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{I}_{f_{\mathcal{D}}(x) \neq y}]$$

$$= \mathbb{E}_{x \sim \mathcal{D}_X} \left[ \mathbb{E}_{y \sim \mathcal{D}_{Y|x}}[\mathbb{I}_{f_{\mathcal{D}}(x) \neq y} | X = x] \right]$$

$$= \mathbb{E}_{x \sim \mathcal{D}_X} [\alpha_x]$$

$$\leq \mathbb{E}_{x \sim \mathcal{D}_X} \left[ \mathbb{E}_{y \sim \mathcal{D}_{Y|x}}[\mathbb{I}_{g(x) \neq y} | X = x] \right]$$

$$= L_{\mathcal{D}}(g) .$$

---

\(^1\)As we shall see, $g$ might be non-deterministic.
8. (a) This was proved in the previous exercise.
   
   (b) We proved in the previous exercise that for every distribution $D$, the bayes optimal predictor $f_D$ is optimal w.r.t. $D$.
   
   (c) Choose any distribution $D$. Then $A$ is not better than $f_D$ w.r.t. $D$.

9. (a) Suppose that $\mathcal{H}$ is PAC learnable in the one-oracle model. Let $A$ be an algorithm which learns $\mathcal{H}$ and denote by $m_H$ the function that determines its sample complexity. We prove that $\mathcal{H}$ is PAC learnable also in the two-oracle model.

   Let $D$ be a distribution over $\mathcal{X} \times \{0, 1\}$. Note that drawing points from the negative and positive oracles with equal provability is equivalent to obtaining i.i.d. examples from a distribution $D'$ which gives equal probability to positive and negative examples. Formally, for every subset $E \subseteq \mathcal{X}$ we have

   $$D'[E] = \frac{1}{2} D^+[E] + \frac{1}{2} D^-[E].$$

   Thus, $D'[\{x : f(x) = 1\}] = D'[\{x : f(x) = 0\}] = \frac{1}{2}$. If we let $A$ an access to a training set which is drawn i.i.d. according to $D'$ with size $m_H(\epsilon/2, \delta)$, then with probability at least $1 - \delta$, $A$ returns $h$ with

   $$\epsilon/2 \geq L_{(D', f)}(h) = \mathbb{P}_{x \sim D'}[h(x) \neq f(x)]$$

   $$= \mathbb{P}_{x \sim D'}[f(x) = 1, h(x) = 0] + \mathbb{P}_{x \sim D'}[f(x) = 0, h(x) = 1]$$

   $$= \mathbb{P}_{x \sim D'}[f(x) = 1] \cdot \mathbb{P}_{x \sim D'}[h(x) = 0 | f(x) = 1]$$

   $$+ \mathbb{P}_{x \sim D'}[f(x) = 0] \cdot \mathbb{P}_{x \sim D'}[h(x) = 1 | f(x) = 0]$$

   $$= \mathbb{P}_{x \sim D'}[f(x) = 1] \cdot \mathbb{P}_{x \sim D'}[h(x) = 0 | f(x) = 1]$$

   $$+ \mathbb{P}_{x \sim D'}[f(x) = 0] \cdot \mathbb{P}_{x \sim D'}[h(x) = 1 | f(x) = 0]$$

   $$= \frac{1}{2} L_{(D^+, f)}(h) + \frac{1}{2} L_{(D^-, f)}(h).$$

   This implies that with probability at least $1 - \delta$, both

   $$L_{(D^+, f)}(h) \leq \epsilon$$

   $$L_{(D^-, f)}(h) \leq \epsilon.$$