A Probability Review

1. The first thing we notice is that the problem is referring to a fixed hypothesis h and therefore

$$R(h) = E[\hat{R}(h)] \tag{1}$$

Now we consider the definition of the variance and substitute equation 1:

$$\begin{aligned}\text{Var}[\hat{R}(h)] &= E[(\hat{R}(h) - E[\hat{R}(h)])^2] \\ &= E[(\hat{R}(h) - R(h))^2]\end{aligned} \tag{2}$$

Using the identity $E[X^2] = \int_0^\infty Pr[X^2 > t]dt$ in equation 2 we get

$$\text{Var}[\hat{R}(h)] = \int_0^\infty Pr[(\hat{R}(h) - R(h))^2 > t]dt \tag{3}$$

Which can be rewritten as follows for any value of $u \in (0, \infty)$:

$$\text{Var}[\hat{R}(h)] = \int_0^u Pr[(\hat{R}(h) - R(h))^2 > t]dt + \int_u^\infty Pr[(\hat{R}(h) - R(h))^2 > t]dt \tag{4}$$

The most basic definition of probability tells us that $P(A) \in [0, 1]$ for any event A, which implies $P(A) \leq 1$ and so we can write

$$\begin{aligned}Pr[(\hat{R}(h) - R(h))^2 > t]dt &\leq 1 \\ \int_0^u Pr[(\hat{R}(h) - R(h))^2 > t]dt &\leq \int_0^u dt \\ &\leq u\end{aligned} \tag{5}$$

Replacing inequality 5 in equation 4 we get

$$\text{Var}[\hat{R}(h)] \leq u + \int_u^\infty Pr[(\hat{R}(h) - R(h))^2 > t]dt \tag{6}$$

Now we refer to the inequality provided in the problem definition and make $\epsilon^2 = t$:

$$\begin{aligned}Pr[|\hat{R}(h) - R(h)| > \epsilon] &\leq 2e^{-2m\epsilon^2} \\ &\iff \\ Pr[(\hat{R}(h) - R(h))^2 > \epsilon^2] &\leq 2e^{-2m\epsilon^2} \\ Pr[(\hat{R}(h) - R(h))^2 > t] &\leq 2e^{-2mt}\end{aligned} \tag{7}$$

Replacing inequality 7 in inequality 6 we get

$$\begin{aligned}\text{Var}[\hat{R}(h)] &\leq u + \int_u^\infty 2e^{-2mt}dt \\ &\leq u - \frac{1}{m}e^{-2mt}\Big|_u^\infty \\ &\leq u + \frac{1}{m}e^{-2mu}\end{aligned} \tag{8}$$

Finally, we find the value of u that minimizes the upper bound for the variance (right-hand side of inequality 8) by setting its derivative equals to zero and solving for u:

$$\begin{aligned}\frac{d}{du}\left(u + \frac{1}{m}e^{-2mu}\right) &= 0 \\ 1 - 2e^{-2mu} &= 0 \\ e^{2mu} &= 2 \\ u &= \frac{\log(2)}{2m}\end{aligned} \tag{9}$$

Replacing equation 9 in inequality 8 we get

$$\begin{aligned}
\text{Var}[\hat{R}(h)] &\leq \frac{\log(2)}{2m} + \frac{1}{m} e^{-2m\left(\frac{\log(2)}{2m}\right)} \\
&\leq \frac{\log(2)}{2m} + \frac{1}{2m} \\
&\leq \frac{\log(2e)}{2m} \quad (Q.E.D.)
\end{aligned}$$

B PAC Learning

1. According to the problem statement, a threshold function f_c is defined as follows:

$$f_c(x) = \begin{cases} 0: & x < c \\ 1: & x \geq c \end{cases}$$

If we consider a sample $S = \{x_i, f_c(x_i)\}$ of size m drawn from a distribution D, we can find a separator $\gamma \in \mathbb{R}$ that divides this set in such a way that for all sample points labeled 0 we have $x_i \leq \gamma$ and for all sample points labeled 1 we have $x_i > \gamma$. We define our learning algorithm L with the following hypothesis h_S based on γ:

$$h_S(x) = \begin{cases} 0: & x < \gamma \\ 1: & x \geq \gamma \end{cases}$$

Now we define G as the set of valid choices for γ i.e., the interval between the rightmost sample point labeled -1 and the leftmost sample point labeled 1. Under this definition, G is a random interval that depends on the sample S, and if it is narrow enough then γ will be very close to the true value of c and our algorithm will have a small error R, since our algorithm can only make mistakes within G.
Now we notice that $R(h_S) = Pr_{x \sim D}[h_S(x) \neq f_c(x)]$ is equivalent to the amount of weight that the distribution D puts in the interval between c and γ, and so we would like to find an upper bound on $Pr[R(h_S) > \epsilon]$.
First we set $\epsilon > 0$ and define c_1 and c_2 as follows:

$$c_1 = \max_{v<c}(Pr_{x \sim D}[v \leq x \leq c] \geq \epsilon)$$
$$c_2 = \min_{v>c}(Pr_{x \sim D}[v \leq x \leq c] \geq \epsilon)$$

If the input sample S contains at least one point in $C_1 = [c_1, c]$ and one point in $C_2 = [c, c_2]$, then our algorithm must output a threshold value $\gamma \in [c_1, c_2]$ and it is easy to see that any such value will have error no more than ϵ with respect to c. Therefore, $Pr[R(h_S) > \epsilon]$ implies that the sample S does not contain a point in at least one of these regions and we write:

$$Pr[R(h_S) > \epsilon] \leq Pr[x_1 \notin C_1 \wedge \cdots \wedge x_n \notin C_1] + Pr[x_1 \notin C_2 \wedge \cdots \wedge x_n \notin C_2]$$

Then we observe that

$$\begin{aligned} Pr[x_1 \notin C_1 \wedge \cdots \wedge x_n \notin C_1] &= \prod_{i=1}^{m} Pr[x_i \notin C_1] \\ &\leq (1-\epsilon)^m \\ &\leq e^{-\epsilon m} \end{aligned}$$

With a similar procedure we can bound the probability that no point in the sample falls in C_2 and therefore we have that

$$Pr[R(h_S) > \epsilon] \leq 2e^{-\epsilon m}$$

To finish the proof we set δ to match the upper bound and solve for m:

$$\begin{aligned} 2e^{-\epsilon m} &\leq \delta \\ m &\geq \frac{1}{\epsilon} \ln\left(\frac{2}{\delta}\right) \end{aligned}$$

With this we can assure that L is a PAC-learning algorithm for C, and therefore for a sample S of size $m \geq \frac{1}{\epsilon} \ln\left(\frac{2}{\delta}\right)$, L will return a hypothesis h_S such that $Pr[R(h_S) \leq \epsilon] \geq 1 - \delta$.

2. Let us consider a concept function f_c of the following form:

$$f_{c_x c_y}(x,y) = \begin{cases} 0: & x < c_x, y < c_y \\ 1: & x \geq c_x, y \geq c_y \end{cases}$$

If we look only at one of the coordinates for the m sample points in S, we can apply our PAC-learning algorithm L from the previous exercise:

$$Pr[R(h_{S_x}) > \epsilon] \leq 2e^{-\epsilon m}$$
$$Pr[R(h_{S_y}) > \epsilon] \leq 2e^{-\epsilon m}$$

Where h_{S_x} and h_{S_y} are the hypotheses returned by L for the x-coordinate and y-coordinate, respectively.
Now we can define our hypothesis h_S:

$$h_S(x,y) = h_{S_x} \wedge h_{S_y}$$

Under this definition, it is easy to see that h_S will only make a mistake when h_{S_x} or h_{S_y} make a mistake. Thinking of it in terms of the probability $Pr[R(h_S) > \epsilon]$ that we are trying to upper bound we write:

$$\begin{aligned} Pr[R(h_S) > \epsilon] &= Pr[R(h_{S_x}) > \epsilon] \vee Pr[R(h_{S_y}) > \epsilon] \\ &= Pr[R(h_{S_x}) > \epsilon] + Pr[R(h_{S_y}) > \epsilon] \\ &\leq 4e^{-\epsilon m} \end{aligned}$$

The first step follows from the fact that x and y are the coordinates of the sample points and therefore random and independent. The second step is a simple substitution that follows from adding the results for h_{S_x} and h_{S_y}.
Now we are ready to finish our proof by setting δ to match the upper bound and solving for m:

$$\begin{aligned} 4e^{-\epsilon m} &\leq \delta \\ m &\geq \frac{1}{\epsilon} \ln\left(\frac{4}{\delta}\right) \end{aligned}$$

To generalize our learning algorithm to other concept functions in C_2, we simply have to keep track of the minimum coordinates of positive and negative sample points in S and define the partial hypotheses accordingly:

Condition x	Condition y	Positive concept $f_{c_x c_y}$	Positive h_{S_x}	Positive h_{S_y}
$\min_x^- < \min_x^+$	$\min_y^- < \min_y^+$	$x \geq c_x, y \geq c_y$	$x \geq \gamma_x$	$y \geq \gamma_y$
$\min_x^- < \min_x^+$	$\min_y^- > \min_y^+$	$x \geq c_x, y \leq c_y$	$x \geq \gamma_x$	$y \leq \gamma_y$
$\min_x^- > \min_x^+$	$\min_y^- < \min_y^+$	$x \leq c_x, y \geq c_y$	$x \leq \gamma_x$	$y \geq \gamma_y$
$\min_x^- > \min_x^+$	$\min_y^- > \min_y^+$	$x \leq c_x, y \leq c_y$	$x \leq \gamma_x$	$y \leq \gamma_y$

Where the notation $\min_x^-$ means the minimum value in the x-coordinate for all points classified as negative in the sample S. With these definitions for the partial hypotheses h_{S_x} and h_{S_y}, we can keep our final hypothesis definition constant for all concepts in C_2: $h_S(x,y) = h_{S_x} \wedge h_{S_y}$.