Name: Krishna Chaitanya Sripada

Ans 2.3

Since the set concepts are of the form c = {(x,y): $x^2 + y^2 \leq r^2$} , the circle is around the origin with radius r. Let us choose a smaller radius q such that both of them have the same center i.e., origin.

Let A denote the region between circle with radius 'r' and circle with radius 'q' such that A = $\{x : q \leq \|x\| \leq r\}$. Let $P_r[A]$ denote the probability mass of the region defined by A, that is the probability that a point randomly drawn falls within A. Since errors made by the PAC-learning algorithm can be only due to points falling inside A, we can assume that $P_r[A] > \varepsilon$; otherwise the error of A is less than or equal to ε regardless of the training sample.

Let R∈c be a target concept and H be a hypothesis and by contraposition, if R(A) > ε, then any point in H chosen accordingly will "miss" region A with a probability of at most 1 - ε. Therefore, we get,

$$\begin{aligned} P_r[R(A) > \varepsilon] &\leq P_r[\{A \cap H = \varnothing\}] \\ &\leq (1-\varepsilon)^m \\ &\leq e^{-m\varepsilon} \end{aligned}$$

where for the last step, the identity 1-x $\leq e^{-x}$ is used which is valid for all x $\in \mathbb{R}$.

For any δ > 0, to ensure that $P_r[R(A) > \varepsilon] \leq \delta$, we can impose,

$$e^{-m\varepsilon} \leq \delta \Leftrightarrow \text{m} \geq (1/\varepsilon)\log(1/\delta)$$

Ans 2.4

Given X = $\mathbb{R}^2$ and the set of concepts are of the form c = $\{x \in \mathbb{R}^2 : \|x - x_0\| \leq r\}$ for some point $x_0 \in \mathbb{R}^2$ and real number r. Also the complexity is $m \geq (3/\varepsilon)\log(3/\delta)$ with three regions r_1, r_2, r_3 drawn around the edge of concept c have probability of $\varepsilon/3$ each.

Gertrude is relying on the implication that generalization $error > \varepsilon \Rightarrow H \cap r_i = \varnothing$ for some i and hypothesis H. Below is the illustration of an example where we have one training point in each region. The points in r_1 and r_2 are very close together, and the point in r_3 is very close to region r_1. In this data, the learned circle includes these points and one diameter approximately traverses the corners of r_1. In the illustration below, the circle with the thick border is our target circle and the darkened areas are the errors of this hypothesis. Apparently, the error can be greater that ε even while $H \cap r_i = \varnothing$ $\forall$ i and this invalidates Gertrude's proof.

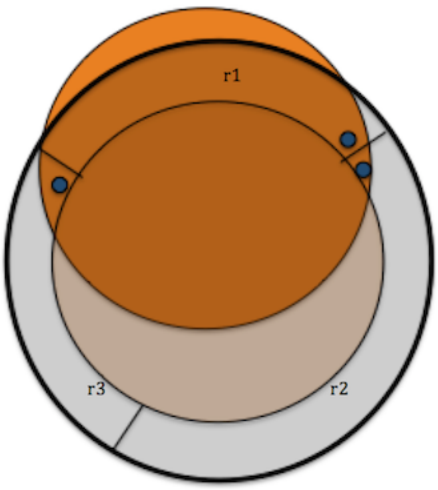

Figure 1 :Non-concentric circles

Ans 2.6

(a) The probability that $R^{'}$ misses region r_j is the product of the probability p for each point x_i of the training sample that
i. Doesn't fall in r_j or be positive.
ii. Fall in r_j with the label flipped to negative because of the noise.
Then, we have,

$$\begin{aligned}
p &= P_r[x \notin r_j \vee (x \in r_j \wedge x \text{ is positive} \wedge \text{label of } x \text{ is flipped})] \\
&= P_r[x \notin r_j \vee (x \in r_j \wedge \text{ label of } x \text{ is flipped})] \\
&= P_r[x \notin r_j] + P_r[(x \in r_j \wedge \text{ label of } x \text{ is flipped})] \\
&= (1 - P_r[x \in r_j]) + \eta P_r[x \in r_j] \\
&= (1-\eta)(1 - P_r[x \notin r_j]) + \eta \\
&\leq (1-\eta)(1-\varepsilon/4) + \eta \text{ (by the definition of PAC learnability)} \\
&= (1-\varepsilon/4) + \eta\varepsilon/4 \\
&\leq 1 - \varepsilon(1-\eta^{'})/4
\end{aligned}$$

(b) The probability that $P_r[R(R^{'}) > \varepsilon]$ is upper bound by the probability that $R^{'}$ misses at least one region r_j. Thus, by union bound, we get,

$$P_r[R(R^{'}) > \varepsilon] \leq 4(1-\varepsilon(1-\eta^{'})/4)^m$$
$$P_r[R(R^{'}) > \varepsilon] \leq 4e^{-m\varepsilon(1-\eta^{'})/4}$$

By setting δ to match the upper bound will result in a probability of at least 1 - δ, $m \geq \frac{4}{(1-\eta^{'})\varepsilon} \log \frac{4}{\delta}$ with $R(R^{'}) \leq \varepsilon$

Ans 3.5

Consider the case where H is reduced to the constant hypothesis $h_1 : x \mapsto 1$ and $h_{-1} : x \mapsto -1$. Then by definition of Rademacher complexity,

$$\hat{R}_s(H) = \frac{1}{m} E_\sigma[sup\{\sum_{i=1}^{m} \sigma_i, \sum_{i=1}^{m} -\sigma_i\}] = \frac{1}{m} E_\sigma[|\sum_{i=1}^{m} \sigma_i|]$$

Let $X = \sum_{i=1}^{m} \sigma_i$. Since E[$X^2$] = $E[\sum_{i,j=1}^{m} \sigma_i\sigma_j]$ and $\forall i \neq j$ and σ_i and σ_j are independent, we get $E[\sigma_i\sigma_j] = E[\sigma_i]E[\sigma_j] = 0$. Thus,

$$E[X^2] = E[\sum_i^m \sigma_i\sigma_i] = E[\sum_i^m \sigma_i^2].$$

Since $m = E[X^2]$, it can be rewritten as $E[|X|^{\frac{2}{3}}|X|^{\frac{4}{3}}] \leq E[|X|]^{\frac{2}{3}} E[X^4]^{\frac{1}{3}}$.

Therefore,

$$E[|X|] \geq \frac{m^{\frac{3}{2}}}{E[X^4]^{\frac{1}{2}}} = \frac{m^{\frac{3}{2}}}{\sqrt{E[\sum_{i=1}^{m} \sigma_i^4 + 3\sum_{i \neq j} \sigma_i^2\sigma_j^2]}} = \frac{m^{\frac{3}{2}}}{\sqrt{m+3m(m-1)}} = \frac{m^{\frac{3}{2}}}{\sqrt{m(3m-2)}} \geq \frac{m^{\frac{3}{2}}}{\sqrt{m(3m)}} = \sqrt{\frac{m}{3}}.$$

Thus,

$$\hat{R}_s(H) \geq \sqrt{\frac{m}{3}}$$

Since R_m(H) $\leq \hat{R}_s(H) + O(\frac{VCdim(H)}{\sqrt{m}})$, it implies R_m(H) $\leq O(\frac{VCdim(H)}{\sqrt{m}})$, which contradicts R_m(H) $\leq O(\frac{VCdim(H)}{m})$.

Ans 3.6

A sequence of $2k+1$ points on a line can't be shattered if successive points are labeled with alternate labels starting with a positive label. We need to choose intervals which contain a longest sequence of consecutive positive sample points and we can have at most $2k$ such intervals. Thus, VC dimension of the class of union of k intervals on a real line is $2k$.

Ans 3.12

(a) For any x $\in \mathbb{R}$, let there exist an ω with labels $--+-$. Then $\sin(\omega x) < 0$, $\sin(2\omega x) < 0$, $\sin(3\omega x) > 0$ and $\sin(4\omega x) < 0$. If we show that this implies $\sin^2(\omega x) < \frac{1}{2}$ and $\sin^2(\omega x) \geq \frac{3}{4}$, then it will be a contradiction.

Using the identity $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$ and since $\sin(4\omega x) < 0$, we have,

$$2\sin(2\omega x)\cos(2\omega x) = \sin(4\omega x) < 0.$$

Since $\sin(2\omega x) < 0$, we can divide both sides of the inequality by $2\sin(2\omega x)$ to get $\cos(2\omega x) > 0$. Applying the identity $\cos(2\theta) = 1 - 2\sin^2(\theta)$ yields $1 - 2\sin^2(\omega x) > 0$, or $\sin^2(\omega x) < \frac{1}{2}$.

Using the identity $\sin(3\theta) = 3\sin(\theta) - 4\sin^3(\theta)$ and $\sin(3\omega x) \geq 0$, we have

$$3\sin(\omega x) - 4\sin^3(\omega x) = \sin(3\omega x) \geq 0.$$

Since $\sin(\omega x) < 0$ we can divide both sides of the inequality by $\sin(\omega x)$ to get $3 - 4\sin^2(\omega x) \leq 0$ or $\sin^2(\omega x) \geq \frac{3}{4}$. Hence we have proved the contraction and thus $\forall x \in \mathbb{R}$, the points $x, 2x, 3x$ and $4x$ cannot be shattered by this family of sine functions.

(b) For any $m > 0$, consider points $(x_1, x_2, ..., x_m)$ with arbitrary labels $(y_1, y_2, ..., y_m) \in \{-1, +1\}^m$. Now, let parameter $\omega = \pi(1 + \sum_{i=1}^{m} 2^i y_i^{'})$ where $y_i^{'} = \frac{1-y_i}{2}$. If we can show that this parameter will classify the entire sample for any $m > 0$ and choice of labels, then we can show that the VC-dimension of the family of sine functions is infinite.

$\forall \in [1, m]$, we have

$$\begin{aligned}\omega x_j = \omega 2^{-j} &= \pi(2^{-j} + \sum_{i=1}^{m} 2^{i-j} y_i^{'}) \\ &= \pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y_i^{'}) + y_j^{'} + (\sum_{i=1}^{m-j} 2^i y_i^{'}))\end{aligned}$$

The last term can be ignored as it only contributes multiples of 2π. Since $y_i^{'} \in \{0, 1\}$ the sum of remaining terms is,
$\pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y_i^{'}) + y_j^{'}) = \pi(\sum_{i=1}^{j-1} 2^{-i} y_i^{'} + 2^{-j} + y_j^{'})$
Now the upper and lower bounds are as follows:

$$\pi(\sum_{i=1}^{j-1} 2^{-i} y_i^{'} + 2^{-j} + y_j^{'}) \leq \pi(\sum_{i=1}^{j} 2^{-i} + y_j^{'}) < \pi(1 + y_j^{'})$$
$$\pi(\sum_{i=1}^{j-1} 2^{-i} y_i^{'} + 2^{-j} + y_j^{'}) > \pi y_j^{'}$$

Thus, if $y_j = 1$ we have $y_j^{'} = 0$ and $0 < \omega x_j < \pi$, which implies $\sin(\omega x_j) = 1$. Similarly, for $y_j = -1$ we have $\sin(\omega x_j) = -1$

Ans 3.19

(a) By definition of Oskar's prediction rule,

$$\begin{aligned}error(f_o) &= P_r[f_o(S) \neq x] \\ &= P_r[f_o(S) = x_A \wedge x = x_B] + P_r[f_o(S) = x_B \wedge x = x_A] \\ &= P_r[N(S) < \tfrac{m}{2} | x = x_B] P_r[x = x_B] + P_r[N(S) \geq \tfrac{m}{2} | x = x_A] P_r[x = x_A] \\ &= \tfrac{1}{2} P_r[NS < \tfrac{m}{2} | x = x_B] + \tfrac{1}{2} P_r[N(S) \geq \tfrac{m}{2} | x = x_A] \geq \tfrac{1}{2} P_r[N(S) \geq \tfrac{m}{2} | x = x_A]\end{aligned}$$

(b) Since $P_r[N(S) \geq \frac{m}{2} | x = x_A] = P_r[B(m, p) \geq k]$, with $p = \frac{1-\varepsilon}{2}, k = \frac{m}{2}$ and $mp \leq k \leq m(1-p)$. Thus, by the binomial inequality given in the appendix,

$$error(f_o) \geq \tfrac{1}{2} P_r[N \geq \frac{\frac{m\varepsilon}{2}}{\sqrt{\frac{1}{4(1-\varepsilon^2)m}}}] = \tfrac{1}{2} P_r[N \geq \varepsilon \sqrt{\tfrac{m}{1-\varepsilon^2}}]$$

Using the second inequality in the appendix, we get,

$$error(f_o) > \tfrac{1}{4}[1 - [1 - e^{-\frac{m\varepsilon^2}{1-\varepsilon^2}}]^{\frac{1}{2}}]$$

(c) If m is odd, $P_r[N(S) \geq \frac{m}{2} | x = x_A] \geq P_r[N(S) \geq \frac{m+1}{2} | x = x_A]$, we use the lower bound,

$$error(f_o) \geq \tfrac{1}{2} P_r[N(S) \geq \tfrac{m+1}{2} | x = x_A]$$

Thus we can use the lower bound expression with $\lceil \frac{m}{2} \rceil$ instead of $\frac{m}{2}$.
Therefore,

$$error(f_o) > \tfrac{1}{4}[1 - [1 - e^{-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2}}]^{\frac{1}{2}}]$$

(d) If $error(f_o)$ is at most δ where $0 < \delta < 1/4$, then $\frac{1}{4}[1 - [1 - e^{-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2}}]^{\frac{1}{2}}] < \delta$. Upon simplification, we get,

$$\begin{aligned}&e^{-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2}} < 1 - (1 - 4\delta)^2 \\ &e^{-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2}} < 4\delta(2 - 4\delta) \\ &e^{-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2}} < 8\delta(1 - 2\delta)\end{aligned}$$

and solving for m, we get,

$$\begin{aligned}&-\frac{2\lceil \frac{m}{2} \rceil \varepsilon^2}{1-\varepsilon^2} < \log(8\delta(1 - 2\delta)) \\ &-\lceil \tfrac{m}{2} \rceil < \tfrac{1-\varepsilon^2}{2\varepsilon^2} \log(8\delta(1 - 2\delta)) \\ &m > 2\left\lceil \tfrac{1-\varepsilon^2}{2\varepsilon^2} \log(8\delta(1 - 2\delta)) \right\rceil\end{aligned}$$

Thus lower bound varies as $\frac{1}{\varepsilon^2}$.

(e) Let f be an arbitrary rule and X_A denote the set of samples for which $f(S) = x_A$ and F_B the complement. Then, by definition of error,

$$
\begin{aligned}
error(f) &= \sum_{S \in X_A} P_r[S \wedge x_B] + \sum_{S \in X_B} P_r[S \wedge x_A] \\
&= \tfrac{1}{2} \sum_{S \in X_A} P_r[S|x_B] + \tfrac{1}{2} \sum_{S \in X_B} P_r[S|x_A] \\
&= \tfrac{1}{2} \sum_{S \in X_A, N(S) < m/2} P_r[S|x_B] + \tfrac{1}{2} \sum_{S \in X_A, N(S) \geq m/2} P_r[S|x_B] + \tfrac{1}{2} \sum_{S \in X_B, N(S) < m/2} P_r[S|x_A] + \tfrac{1}{2} \sum_{S \in X_B, N(S) \geq m/2} P_r[S|x_A]
\end{aligned}
$$

If N(S) $\geq$ m/2, then $P_r[S|x_B] \geq P_r[S|x_A]$ and if N(S) < m/2, then $P_r[S|x_A] \geq P_r[S|x_B]$. Thus the lower bound is,

$$
\begin{aligned}
error(f) &\geq \tfrac{1}{2} \sum_{S \in X_A, N(S) < m/2} P_r[S|x_B] + \tfrac{1}{2} \sum_{S \in X_A, N(S) \geq m/2} P_r[S|x_A] + \tfrac{1}{2} \sum_{S \in X_B, N(S) < m/2} P_r[S|x_B] + \tfrac{1}{2} \sum_{S \in X_B, N(S) \geq m/2} P_r[S|x_A] \\
&= \tfrac{1}{2} \sum_{N(S) < m/2} P_r[S|x_B] + \tfrac{1}{2} \sum_{N(S) \geq m/2} P_r[S|x_A] \\
&= error(f_o)
\end{aligned}
$$

Thus we can conclude that the lower bound can be applied to all rules.