HW-5

Q1 (a) $P = \binom{N}{k} \mu^k (1-\mu)^{N-k}$ $\nu = \frac{k}{N}$

(i) $N = 10$

$\mu = 0.05$

If $\nu = \frac{k}{N} = 0 \Rightarrow k = 0$

$\therefore P(\nu = 0 \mid N, \mu) = P(k=0 \mid 10, 0.05)$

$= \binom{10}{0} (0.05)^0 (1-0.05)^{10}$

$= (0.95)^{10}$

$\therefore P(\nu \neq 0 \mid N, \mu) = 1 - P(k=0 \mid 10, 0.05)$

$= 1 - 0.95^{10}$

- For 1 coin, $P(\nu = 0 \mid N, \mu) = 0.95^{10}$
- For 1000 coins,

$P(\text{at least one } \nu_i = 0 \mid N, \mu)$

$= 1 - P(\Pi^{10} \nu_i = 0 \mid N, \mu)$

$= 1 - \prod_{i=1}^{1000} P(\nu_i = 0 \mid N, \mu)$

$= 1 - (1 - (0.95)^{10})^{1000}$

- For 1000000 coins, same as before,

$P(\text{at least one } \nu_i = 0 \mid N, \mu)$

$= 1 - (1 - (0.95)^{10})^{1000000}$

(ii) $\mu = 0.8$

$$P(\nu = 0 \mid N, \mu) = \binom{10}{0} (0.8)^0 (1 - 0.8)^{10}$$

$$= (0.2)^{10}$$

$$P(\nu \neq 0 \mid N, \mu) = 1 - 0.2^{10}$$

For 1 coin, $P = 0.2^{10}$

For 1000 coins, $P = 1 - (1 - 0.2^{10})^{1000}$

For 1000000 coins, $P = 1 - (1 - 0.2^{10})^{1000000}$

(b) (i) 1000 coins represent the training dataset $\mathcal{D}_{train}$

(ii) The calculation in part (a) is for $E_{in}(h)$

(iii) $\nu = E_{out}(h) = 0.05$

$\mu = E_{in}(h) = 1 - (1 - 0.95^{10})^{1000}$

$M = 1$

Q2. $\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$

(a) $M = 1$ $\delta = 0.03$ $\epsilon \leq 0.05$

So,

$$\sqrt{\frac{1}{2N} \ln \frac{2}{0.03}} \leq 0.05$$

Solving for N,

$$N \geq 840$$

(b) $M = 100$ $\delta = 0.03$ $\epsilon \leq 0.05$

So,

$$\sqrt{\frac{1}{2N} \ln \frac{200}{0.03}} \leq 0.05$$

Solving for N,

$$N \geq 1761$$

(c) $M = 10000$ $\delta = 0.03$ $\epsilon \leq 0.05$

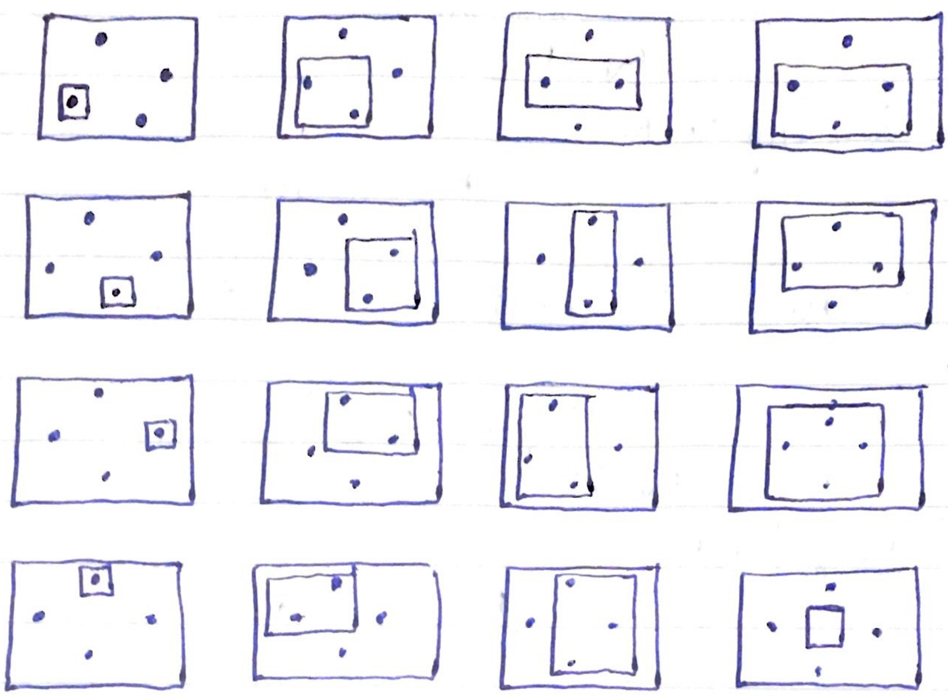
So,

$$\sqrt{\frac{1}{2N} \ln \frac{20000}{0.03}} \leq 0.05$$

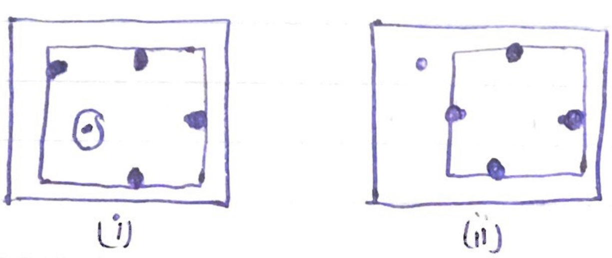
Solving for N,

$$N \geq 2682$$

Q3. The learning model is implying $\mathcal{X} = \mathbb{R}^2$.
Now, if we consider four points, we can show that hypothesis set of positive rectangle is capable of shattering it due to four degrees of freedom.



To prove that the positive set of rectangles can not shatter five points, we can use as an example:



(i) (ii)

If four points are touching as shown in (ii), it is impossible to make the dichotomy where the outer point is enveloped, but the point on the adjacent side is not as shown in (i). Thus, it is not possible to shatter a set of five points.

Thus, $m_H(4) = 2^4$

$m_H(5) < 2^5$

Q4.

(a) The expectation operator $\mathbb{E}$ is a linear operator. So, if any linear combination of hypothesis in H is also a hypothesis in H, i.e., H is closed under linear combination, then $\bar{g} \in \mathcal{H}$.

We know, $$E(X) = \frac{\sum_{j=1}^{n} x_j f_j}{\sum_{j=1}^{n} f_j}$$

which is linear.

Also, $$\bar{g}(x) = \frac{1}{K} \sum_{k=1}^{K} g_k(x)$$

And if $h_1, h_2, \dots, h_k \in H$

Then $\alpha_1 h_1 + \alpha_2 h_2 + \dots + \alpha_k h_k \in H$

(under linear combination)

And $$\bar{g}(x) = \frac{1}{k}\left[g_1(x) + g_2(x) + \dots + g_k(x)\right]$$

where $\alpha_1 = \alpha_2 = \dots = \alpha_k = \frac{1}{k}$

Hence, $\bar{g} \in H$.

(b). Let $X = Y = \mathbb{R}$

$H = 1, 0$

Then, unless one of the hypothesis occurs with probability 0, $\bar{g}$ is not in the model's hypothesis set H.

(c) For binary classification, $\bar{g}$ is <u>not</u> a binary function.

Q5 (a) If we assume H is fixed and we increase the complexity of f, deterministic noise in general will go up because it is harder for any hypothesis H to approximate f. This causes the bias and variance components of expected out-of sample error to go up.
If expected in sample error is low, then model is overfitting and if it is high, then model is underfitting. Generally, there is a higher tendency to overfit.

(b) If we assume f is fixed and we decrease the complexity of H, deterministic noise will go up generally as H has lower chance to approximate f. Overfitting goes up. However, as H is less complex, the model has less variance and less tendency to overfit, so overfitting will go down. Based on these two competing terms, often the second term wins and overfitting, in general, will go down.