Homework 3

Problem 1

a.

i) Proving union bound

We know $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Since probabilities are non-negative, $P(A \cap B) \geq 0$

$\therefore P(A \cup B) \leq P(A) + P(B)$

When is the union bound tight?

- it becomes tight when A and B are disjoint events $(A \cap B = \emptyset)$
- $P(A \cap B) = 0$ which means $P(A \cup B) = P(A) + P(B)$

ii) Proving union bound for multiple events

The probability of the union is

$$P(\mathbb{U}_{p=1} A_p) = \sum_{p=1} P(A_p) - \sum_{a \leq i < j \leq P} P(A_i \cap A_j) + \cdots \pm P(A_1 \cap \cdots \cap A_p)$$

However, the intersections will all be ≥ 0 so it can be rewritten as:

$$P(\mathbb{U}_{p=1} A_p) = \sum_{p=1} P(A_p)$$

$\therefore$ we have proven union bound for P events

b.

i) Union Bound
$P(A \cup B) \leq P(A) + P(B)$ where $P(A) = P(B) = \frac{1}{6}$

$$\therefore P(A \cup B) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

ii) Exact Probability
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where $P(A) = P(B) = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$

$$= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$$

Gap? The difference between part i) and ii) is $\frac{1}{36}$ which is the amount that union bound overestimates the exact probability.

c.

H_{bol} consists of all Boolean functions mapping $\{0,1\}^k$ to {0,1} so there are 2^k possible inputs. This means:

$$|H_{bol}| = 2^{2^k}$$

We can substitute this into the bound.

$$N \geq \frac{1}{\epsilon}\log\left(\frac{2^{2^k}}{\delta}\right)$$

The logarithm simplifies to:

$$\log\left(\frac{2^{2^k}}{\delta}\right) = \log\left(\frac{1}{\delta}\right) = 2^k \log 2 + \log\left(\frac{1}{\delta}\right)$$

with $\log_2 = 1$

$$\log\left(\frac{2^{2^k}}{\delta}\right) = 2^k + \log\left(\frac{1}{\delta}\right)$$

Therefore, we can say that the required size of the training set is:

$$N \geq \frac{1}{\epsilon}(2^k + \log\left(\frac{1}{\delta}\right))$$

d.

i) Risk $R(h)$?
h always guesses tails

- risk $R(h)$ is probability that h makes incorrect predictions
- $\therefore R(h) = p$
  ii)

$$P\left[|R(h) - \hat{R}(h)| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N}}\right] \geq 1 - \delta$$

To ensure $|R(h) - \hat{R}(h)| \leq 0.03$, the confidence interval $\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N}}$ must be at most 0.03. We will Rearrange for N

$$\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N}} \leq 0.03$$

$$N \geq \frac{\log\left(\frac{2}{\delta}\right)}{2 * 0.03^2}$$

with $\delta = 0.02$ the equation get simplified to

$$N \geq \frac{4.605}{0.0018} = 2558$$

e.

i) The givens from the equation are

- $H = 2500$
- $N = 200$
- $\delta = 0.05$
- $\hat{R}(h) = 0$ because it is the empirical risk over N samples and they have yet to have an incorrect prediction

The equation itself is

$$P(|R(h) - \bar{R}(h)| \leq \sqrt{\frac{\log\left(\frac{2|H|}{\delta}\right)}{2N}} \forall h \in \mathcal{H}) \geq 1 - \delta$$

Filling in the givens, we get:

$$P(|R(h)| \leq \sqrt{\frac{\log\left(\frac{2|2500|}{0.05}\right)}{2(200)}}) \geq 0.95$$

$$R(h) \leq \sqrt{\frac{11.51}{400}} = 0.17$$

$\therefore$ we can say with 95% confidence there is roughly a 17% chance the selected person predicts incorrectly the next bill

ii)

Same givens from i) but this time $\hat{R}(h) = \frac{20}{200} = .1$ because they have predicted 20 bills incorrectly.

Filling in the givens, we get:

$$\hat{R}(h) - \sqrt{\frac{\log\left(\frac{2|2500|}{0.05}\right)}{2(200)}} \leq R(h) \leq \hat{R}(h) + \sqrt{\frac{\log\left(\frac{2|2500|}{0.05}\right)}{2(200)}}$$

$$0.1 - 0.17 \leq R(h) \leq 0.1 + 0.17$$

The probability can't be ≤ 0 so we are left with

$$0 \leq R(h) \leq 0.27$$

$\therefore$ we can say with 95% confidence there is roughly a 27% chance the selected person predicts incorrectly the next bill

Problem 2

a.

i) nonnegativity

$RC_S(\ell \circ \mathcal{H}) = \frac{1}{N} E_{r\ iid} Rad \sup_{h \in \mathcal{H}} \langle r || (\ell \circ h)[S] \rangle$ is the Empirical Rademacher Complexity

- we need to show: $RC_S(\ell \circ \mathcal{H}) \geq_0$

Properties:

- Supremum: $\sup_{h \in \mathcal{H}} \langle r || (\ell \circ h)[S] \rangle$ that is at least the value of $\langle r || (\ell \circ h)[S] \rangle$ for some $h_0 \in \mathcal{H}$
- Jensen's Inequality: $f(E[v] \leq E[f(v))]$

Convexity:

- The supremum operation over a set is convex as...
- For any $x_1, x_2 \in \mathbb{R}^2$ and $\alpha \in [0, 1]$

$$\sup_{h \in \mathcal{H}}(\alpha x_1 + (1-\alpha)x_2) \leq \alpha \sup_{h \in \mathcal{H}} x_1 + (1-\alpha) \sup_{h \in \mathcal{H}} x_2$$

The function to be applied is

$$f(r) = \sup_{h \in \mathcal{H}} \langle r || (\ell \circ h)[S] \rangle$$

$$\therefore \sup_{h \in \mathcal{H}} \langle E[r] || (\ell \circ h)[S] \rangle \leq E \sup_{h \in \mathcal{H}} \langle r || (\ell \circ h)[S] \rangle$$

The Rademacher variables r_i have zero mean ($E[r_i] = 0$), so

$$\sup_{h \in \mathcal{H}} \langle E[r] || (\ell \circ h)[S] \rangle = \sup_{h \in \mathcal{H}} \langle 0 || (\ell \circ h)[S] \rangle$$

$$\therefore 0 \leq E \sup_{h \in \mathcal{H}} \langle r || (\ell \circ h)[S] \rangle$$

By this we can conclude:

$$RC_S(\ell \circ \mathcal{H}) \geq 0$$

ii) Monotonicity: show $\hat{RC}_s(\ell \circ \mathcal{H}_1) \leq \hat{RC}_s(\ell \circ \mathcal{H}_2)$

We can say:

$$\sup_{h \in \mathcal{H}_1} \langle r || (\ell \circ h)[S] \rangle \leq \sup_{h \in \mathcal{H}_2} \langle r || (\ell \circ h)[S] \rangle$$

as

$$\sup_{b \in B} f(b) \leq \sup_{a \in A} f(a)$$

After taking the expectation over $r \sim iid Rad$, we get

$$E_r \sup_{h \in \mathcal{H}_1} \langle r||(\ell \circ h)[S]\rangle \leq E_r \sup_{h \in \mathcal{H}_2} \langle r||(\ell \circ h)[S]\rangle$$

Adding scale factor of $\frac{1}{N}$

$$\hat{RC}_s(\ell \circ \mathcal{H}_1) = \frac{1}{N} E_r \sup_{h \in \mathcal{H}_1} \langle r||(\ell \circ h)[S]\rangle \leq \hat{RC}_s(\ell \circ \mathcal{H}_2) = \frac{1}{N} E_r \sup_{h \in \mathcal{H}_2} \langle r||(\ell \circ h)[S]\rangle$$

This proves monotonicity

iii) Summation: $RC_S(\ell \circ (\mathcal{H}_1 + \mathcal{H}_2)) = RC_S(\ell \circ \mathcal{H}_1) + RC_S(\ell \circ \mathcal{H}_2)$

$$RC_S(\ell \circ (\mathcal{H}_1 + \mathcal{H}_2) = \frac{1}{N} E_{r \sim iidRad} \sup_{h \in \mathcal{H}_1, h \in \mathcal{H}_2} \langle r||(\ell \circ h_1 + \ell \circ h_2)[S]\rangle$$

This $(\ell \circ h_1 + \ell \circ h_2)[S]$ expands to:

$$[\ell(h_1, z_1) + \ell(h_2, z_2), \ldots, \ell(h_1, z_N) + \ell(h_2, z_N)]$$

Which means $\langle r||(\ell \circ h_1 + \ell \circ h_2)[S]\rangle$ is:

$$= \langle r||(\ell \circ h_1)[S]\rangle + \langle r||(\ell \circ h_2)[S]\rangle$$

The supremum decomposition also holds as the terms $\langle r||(\ell \circ h_1)[S]\rangle$ and $\langle r||(\ell \circ h_2)[S]\rangle$ are independent.

$$= \sup_{h_1 \in \mathcal{H}_1} \langle r||(\ell \circ h_1)[S]\rangle + \sup_{h_2 \in \mathcal{H}_2} \langle r||(\ell \circ h_2)[S]\rangle$$

We can next apply the expecation over $r \sim \text{iid Rad}$

$$RC_S(\ell \circ (\mathcal{H}_1 + \mathcal{H}_2) = E_r[\sup_{h_1 \in \mathcal{H}_1} \langle r||(\ell \circ h_1)[S]\rangle + \sup_{h_2 \in \mathcal{H}_2} \langle r||(\ell \circ h_2)[S]\rangle]$$

Because expectation can be applied linearly...

$$= E_r \sup_{h_1 \in \mathcal{H}_1} \langle r||(\ell \circ h_1)[S]\rangle + E_r \sup_{h_2 \in \mathcal{H}_2} \langle r||(\ell \circ h_2)[S]\rangle$$

After adding the scaling factor, $\frac{1}{N}$, we have proven the property

$$RC_S(\ell \circ (\mathcal{H}_1 + \mathcal{H}_2) = \frac{1}{N} E_r \sup_{h_1 \in \mathcal{H}_1} \langle r||(\ell \circ h_1)[S]\rangle + \frac{1}{N} E_r \sup_{h_2 \in \mathcal{H}_2} \langle r||(\ell \circ h_2)[S]\rangle$$

$$RC_S(\ell \circ (\mathcal{H}_1 + \mathcal{H}_2) = RC_S(\ell \circ \mathcal{H}_1) + RC_S(\ell \circ \mathcal{H}_2)$$

iv) affine transform: show $RC_S(\alpha(\ell \circ \mathcal{H}) + b) = |a| * RC_S(\ell \circ \mathcal{H})$

First we can expand the inner product of the empirical rademacher complexity

$$\langle r, \alpha(\ell \circ h)[S] + b\rangle = \langle r, \alpha(\ell \circ h)[S]\rangle + \langle r, b\rangle$$

Evaluating $\langle r, b\rangle$

$$= b \sum r_i$$

r_i is symmetric as $(r_i \sim \{-1, 1\})$ which means $E[r_i] = 0$

$$\therefore E[\langle r, b\rangle] = b * E_r\left[\sum r_i\right] = b\sum E_r[r_i] = 0$$

This shows b has no impact on the Rademacher complexity and can be dropped

Looking at α, it can be factored out

$$\langle r, \alpha(\ell \circ h)[S]\rangle = \alpha\langle r, (\ell \circ h)[S]\rangle$$

$$\therefore RC_S(\alpha(\ell \circ \mathcal{H}) + b) = \frac{1}{N}E_{r\sim \text{iid Rad}} \sup_{h\in\mathcal{H}} \alpha\langle r, (\ell \circ h)[S]\rangle$$

Yet, we still need to consider the absolute value of α:

- If $\alpha > 0$, the supremum is the same
- If $\alpha < 0$, it is negated. However, since it is taken over all $h \in \mathcal{H}$, the negative is absorbed without loss of generality

$$\therefore \sup_{h\in\mathcal{H}} \alpha\langle r, (\ell \circ h)[S]\rangle = |\alpha| \sup_{h\in\mathcal{H}}\langle r, (\ell \circ h)[S]\rangle$$

This proves the affine transform as it can be substituted back in to show

$$RC_S(\alpha(\ell \circ \mathcal{H}) + b) = |a|\frac{1}{N}E_{r\sim \text{iid Rad}} \sup_{h\in\mathcal{H}}\langle r, (\ell \circ h)[S]\rangle = |a| * RC_S(\ell \circ \mathcal{H})$$

v) Talagrand's contraction lemma
First lets apply the sigmoid function $\sigma(z) = \frac{1}{1+e^z}$ to the Lipschitz Property

$$\sigma'(z) = \frac{d}{dz}\left(\frac{1}{1+e^z}\right) = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)(1-\sigma(z))$$

Next we can maximize the function

$$\sigma''(z) = 1 - 2\sigma(z)$$

Setting $\sigma''(z) = 0$, we get $\sigma(z) = \frac{1}{2}$. Plugging this back into $\sigma'(z)$, we get $\sigma'(z) = \frac{1}{4}$

We can then say: $|\sigma'(z)| \leq \frac{1}{4}$ for all $z \in \mathbb{R}$

$$\therefore |\sigma(z_1) - \sigma(z_2)| \leq \frac{1}{4}|z_1 - z_2|, \forall z_1, z_2 \in R$$

Next we can derive an upper bound of $R\hat{C}_S(\mathcal{H}_1)$ in terms of $R\hat{C}_S(\mathcal{H}_2)$

$$RC_S(\mathcal{H}_1) \leq \frac{1}{4}RC_S(\mathcal{H}_2)$$

In conclusion, the rademacher complexity of $\mathcal{H}_1$ is bounded above by $\frac{1}{4}RC_S(\mathcal{H}_2)$

b.

We are looking to find the empirical gaussian complexity of $\mathcal{H}_{\ell_2}$

$$\langle Xw, g\rangle = w^T X^T g$$

$$\therefore \sup_{\|w\|_2 \leq 1} \langle Xw, g \rangle = \sup_{\|w\|_2 \leq 1} w^T X^T g$$

The supremum becomes...

$$\sup_{\|w\|_2 \leq 1} \|X^T g\|_2$$

after substituting $w = \frac{X^T g}{\|X^T g\|_2}$, which is the normalized version of $X^T g$

From there we can reformulate the gaussian complexity

$$= \frac{1}{d} E_{g \sim N(0,1)} \|X^T g\|_2$$

Using Jensen's Inequality, we can say

$$\frac{1}{d} E\|X^T g\|_2^2 \leq \frac{1}{d} (E\|X^T g\|_2^2)^{1/2}$$

$E\|X^T g\|_2^2$ can be computed further using linearity of expectation

$$= \frac{1}{d} \sum_{j=1}^{d} E\left[\left(\sum_{i=1}^{N} X_{ij} g_i\right)^2\right]$$

$$E\left[\left(\sum_{i=1}^{N} X_{ij} g_i\right)^2\right] = \sum_{i=1}^{N} X_{ij}^2 E[g_2^2] + \sum_{i \neq k} X_{ij} X_{kj} E[g_i g_k]$$

Since g_i are independent and $E[g_i g_k] = 0$ for $i \neq k$, only the diagonal terms remian:

$$= \sum_{i=1}^{N} X_{ij}^2$$

$$\therefore E\|X^T g\|_2^2 = \sum_{j=1}^{d} \sum_{i=1}^{N} X_{ij}^2 = \|X\|_F^2$$

where $\|X\|_F^2$ is the Frobenius nrom of X. Substituting back in we get

$$\frac{1}{d} E\|X^T g\|_2 \leq \frac{1}{d} (\|X\|_F^2)^{1/2}$$

We can then say that the empirical Gaussian complexity of $\mathcal{H}_{\ell_2}$ is upper-bounded by

$$RC_G(\mathcal{H}_{\ell_2}) \leq \frac{\|X\|_F}{d}$$

c.

We can write teh gaussian complexity as

$$= \frac{1}{d} E_{g \sim N(0,1)} \sup_{\|w\|_\infty \leq 1} \sum_{i=1}^{N} g_i \langle w, x_i \rangle$$

Subing in $\langle w, x_i \rangle = \sum_{j=1}^{d} w_j x_{i,j}$

$$\sup_{\|w\|_\infty \le 1} \sum_{i=1}^{N} g_i \langle w, x_i \rangle = \sup_{\|w\|_\infty \le 1} \sum_{j=1}^{d} w_j \sum_{i=1}^{N} g_i x_{i,j}$$

Let $v_j = \sum_{i=1}^{N} g_i x_{i,j}$ with $v = (v_1, v_2 \dots)$

$$= \sup_{\|w\|_\infty \le 1} \langle w, v \rangle$$

Implement bounding using $\|w\|_\infty$:
The supremum $\langle w, v \rangle$ is maximized when $w_j = \text{sign}(v_j)$

$$= \|v\|_1 = \sum_{j=1}^{d} |v_j|$$

$$\therefore G(H_{\ell_\infty}) = \frac{1}{d} E_{g\sim N(0,1)} \|v\|_1$$

$$E\|v\|_1 = \sum_{j=1}^{d} E\|v_j\|$$

where

$$E|v_j| \le \sqrt{Var(v_j)} = \sqrt{\sum^{N} x_{i,j}^2}$$

Additionally, using $\|x_i\|_\infty$

$$\sum_{j=1}^{d} \sum^{N} x_{i,j}^2 \le N \max_i \|x_i\|_\infty^2$$

so

$$E\|v\|_1 \le \sqrt{N} \max_i \|x_i\|_\infty$$

Plugging that back into the main equation we are left with the upper bound being

$$G(H_{\ell_\infty}) = \frac{1}{d} \sqrt{N} \max_i \|x_i\|_\infty$$

Problem 3

a.

Finding the growth function $\prod_{H_{DS}^1}(N)$

- For N = 1, any single point can be labeled as 1 or -1
  - $\prod_{H_{DS}^1}(N) = 2$

- For $N = 2$, consider two points $x_1 < x_2$. There can be four different assignment outcomes. (+1, +1), (-1, +1), (1, -1), (-1,-1)
  - $\prod_{H^1_{DS}}(N) = 4$
- For $N > 2$, H^1_{DS} can no longer realize all possible dichotomies as it can't shatter sets greater than 2

$$\therefore \prod_{H^1_{DS}}(N) = \begin{cases} 2^N & N \leq 2 \\ < 2^N & N > 2 \end{cases}$$

Showing $VCdim(H^1_{DS}) = 2$

- Existence of a set size 2 can be shattered. Let x_1, x_2 be two points were $x_1 < x_2$. There are 4 possible labels
  - all of these can be shattered (+1, +1), (-1, +1), (1, -1), (-1,-1)
- A set of size 3 with x_1, x_2, x_3 where $x_1 < x_2 < x_3$ can't as θ can only create at most 3 distinct regions.

$$\therefore VCdim(H^1_{DS}) = 2$$

b.

Show $VCdim(H) \leq \lfloor \log_2 |H| \rfloor$

If H can shatter a set, S, of size d, the H must assign a distinct hypothesis for every labels of S (2^d possibilities)

$$|H| \geq 2^d$$

$$d \leq \log_2 |H|$$

d is a whole number so...

$$d \leq \lfloor \log_2 |H| \rfloor$$

$$\therefore VCdim(H) \leq \lfloor \log_2 |H| \rfloor$$

c.

i) Show $\prod_C(N) \leq \prod_A(N) + \prod_B(N)$
$C = A \cup B$, aka each hypothesis in C belongs to either A or B. By that logic any dataset S of size N has distinct dichotomies produced by C which is the union of the dichotomies from A or B. Therefore, the most it distinct dichotomies is the sum of the dichotomies which assumes no overlap.
ii)
Using Sauer's Lemma for A and B:

$$\prod_A(N) \leq \sum_{i=0}^{d_A} \binom{N}{i}, \prod_B(N) \leq \sum_{i=0}^{d_B} \binom{N}{i}$$

$$\prod_{C}(N) \leq \sum_{i=0}^{d_A} \binom{N}{i} + \sum_{i=0}^{d_B} \binom{N}{i}$$

For $N \geq d_A + d_B + 2$, note that:

- the combined summartion is still much smaller than 2^N when N becomes large
- we can then say $2^N \geq \sum_{i=0}^{d_A} \binom{N}{i} + \sum_{i=0}^{d_B} \binom{N}{i}$

This proves $\prod_C(N) < 2^N$ for $N \geq d_A + d_B + 2$

Upper bound on $VCdim(C)$:

- it is the largest N such that $\prod_C(N) = 2^N$. Therefore we can say the upper bound is...

$$VCdim(C) \leq d_A + d_B + 1$$

d.

Upper bound of $\hat{RC}_S(H)$

We can apply the $VCdim(H_{HC}) = d + 1$ rule for $N \geq d + 1$:

$$\prod_{H_{HC}}(N) \leq \left(\frac{eN}{d+1}\right)^{d+1}$$

This can be subbed into the following equation

$$\hat{R}_s(H) \leq \sqrt{\frac{2\log_2\left[\left(\frac{eN}{d+1}\right)^{d+1}\right]}{N}}$$

$$\log_2\left[\left(\frac{eN}{d+1}\right)^{d+1}\right] = (d+1)\log_2\left(\frac{eN}{d+1}\right)$$

$$\hat{R}_s(H) \leq \sqrt{\frac{2(d+1)\log_2\left[\left(\frac{eN}{d+1}\right)\right]}{N}}$$

This term can be reduced because N will grow at a rate much faster

$$\hat{R}_s(H) \leq \sqrt{\frac{2(d+1)\log_2 N}{N}}$$