Pregunta 1

(a) De la definición de complejidad de Rademacher tenemos que

$$\mathcal{R}(\mathcal{L}_2, S) = \mathbb{E}\left[\sup_{h\in\mathcal{L}_2} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)\right]$$

Considerando la parametrización de $\mathcal{L}_2$ (y recordando que $(w, b) \in \mathbb{R}^d \times \mathcal{R}$ lo cual obviaré para no cargar la notación) tenemos que

$$\mathcal{R}(\mathcal{L}_2, S) = \mathbb{E}\left[\sup_{\|w\|_2\leq R_2, |b|\leq r} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (w^T X_i + b)\right] = \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_2\leq R, |b|\leq r} \sum_{i=1}^{n} \sigma_i w^T X_i + \sum_{i=1}^{n} \sigma_i b\right]$$

Pero entonces como tenemos el supremo sobre ambos, podemos buscar el supremo sobre cada uno y por estar desagrupado debe ser una cota superior de la búsqueda agrupada. Luego, podemos usar la linealidad de la esperanza

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_2\leq R_2} \sum_{i=1}^{n} \sigma_i w^T X_i + \sup_{|b|\leq r} \sum_{i=1}^{n} \sigma_i b\right] = \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_2\leq R_2} \sum_{i=1}^{n} \sigma_i w^T X_i\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} \sum_{i=1}^{n} \sigma_i b\right]$$

Podemos factorizar lo que no depende de i

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_2\leq R_2} w^T \sum_{i=1}^{n} \sigma_i X_i\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} b \sum_{i=1}^{n} \sigma_i\right]$$

Usando la desigualdad de Cauchy-Schwarz se tiene que

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_2\leq R_2} \|w\|_2 \cdot \left\|\sum_{i=1}^{n} \sigma_i X_i\right\|_2\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} |b| \left|\sum_{i=1}^{n} \sigma_i\right|\right]$$

Y como la norma está acotada resulta en

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{1}{n}\mathbb{E}\left[R_2 \cdot \left\|\sum_{i=1}^{n} \sigma_i X_i\right\|_2\right] + \frac{1}{n}\mathbb{E}\left[r \cdot \left|\sum_{i=1}^{n} \sigma_i\right|\right]$$

Podemos sacar la constante de la esperanza

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \sigma_i X_i\right\|_2\right] + \frac{r}{n}\mathbb{E}\left[\left|\sum_{i=1}^{n} \sigma_i\right|\right]$$

Podemos hacer más fácil de trabajar la expresión si agregamos una raíz cuadrada

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\mathbb{E}\left[\sqrt{\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2^2}\right] + \frac{r}{n}\mathbb{E}\left[\sqrt{\left(\sum_{i=1}^{n}\sigma_i\right)^2}\right]$$

Pero por ser $\sqrt{\cdot}$ una función cóncava, usando la desigualdad de Jensen se tiene que

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{n}\sigma_i X_i\right\|_2^2\right]} + \frac{r}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{n}\sigma_i\right)^2\right]}$$

Separamos la norma

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\mathbb{E}\left[\sum_{k=1}^{d}\left(\sum_{i=1}^{n}\sigma_i X_i^k\right)^2\right]} + \frac{r}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{n}\sigma_i\right)^2\right]}$$

Resolvemos el cuadrado

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\mathbb{E}\left[\sum_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j X_i^k X_j^k\right]} + \frac{r}{n}\sqrt{\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_i\sigma_j\right]}$$

Ahora podemos usar la linealidad de la esperanza

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\sum_{k=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\sigma_i\sigma_j\right]X_i^k X_j^k} + \frac{r}{n}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\sigma_i\sigma_j\right]}$$

Pero si $i \neq j$, entonces $\mathbb{E}[\sigma_i\sigma_j] = \mathbb{E}[\sigma_i]\cdot\mathbb{E}[\sigma_j] = 0\cdot 0 = 0$, debido a la independencia de σ, por lo que solo sobrevive el término de la diagonal

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\sum_{k=1}^{d}\sum_{i=1}^{n}\mathbb{E}\left[\sigma_i^2\right]\left(X_i^k\right)^2} + \frac{r}{n}\sqrt{\sum_{i=1}^{n}\mathbb{E}\left[\sigma_i^2\right]}$$

Pero $\sigma_i^2 = 1$, por lo que

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\sum_{k=1}^{d}\sum_{i=1}^{n}\left(X_i^k\right)^2} + \frac{r}{n}\sqrt{\sum_{i=1}^{n}1}$$

Reordenando el primer término y resolviendo el segundo

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R}{n}\sqrt{\sum_{i=1}^{n}\|X_i\|_2^2} + \frac{r}{n}\sqrt{n}$$

Usando la cota sobre X_i

$$\mathcal{R}(\mathcal{L}_2, S) \leq \frac{R_2}{n}\sqrt{\sum_{i=1}^{n} L_2^2} + \frac{r}{n}\sqrt{n} = \frac{R_2}{n}\sqrt{n \cdot L_2^2} + \frac{r}{n}\sqrt{n} = \frac{L_2 R_2}{\sqrt{n}} + \frac{r}{\sqrt{n}}$$

Tal como queríamos demostrar

(b) De la definición de complejidad de Rademacher tenemos que

$$\mathcal{R}(\mathcal{L}_1, S) = \mathbb{E}\left[\sup_{h\in\mathcal{L}_1} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)\right]$$

Considerando la parametrización de $\mathcal{L}_1$ (y recordando que $(w, b) \in \mathbb{R}^d \times \mathcal{R}$ lo cual obviaré para no cargar la notación) tenemos que

$$\mathcal{R}(\mathcal{L}_1, S) = \mathbb{E}\left[\sup_{\|w\|_1\leq R_1, |b|\leq r} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (w^T X_i + b)\right] = \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_1\leq R_1, |b|\leq r} \sum_{i=1}^{n} \sigma_i w^T X_i + \sum_{i=1}^{n} \sigma_i b\right]$$

Pero entonces como tenemos el supremo sobre ambos, podemos buscar el supremo sobre cada uno y por estar desagrupado debe ser una cota superior de la búsqueda agrupada. Luego, podemos usar la linealidad de la esperanza

$$\mathcal{R}(\mathcal{L}_1, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_1\leq R_1} \sum_{i=1}^{n} \sigma_i w^T X_i + \sup_{|b|\leq r} \sum_{i=1}^{n} \sigma_i b\right] = \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_1\leq R_1} \sum_{i=1}^{n} \sigma_i w^T X_i\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} \sum_{i=1}^{n} \sigma_i b\right]$$

Podemos factorizar lo que no depende de i

$$\mathcal{R}(\mathcal{L}_1, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_1\leq R_1} w^T \sum_{i=1}^{n} \sigma_i X_i\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} b \sum_{i=1}^{n} \sigma_i\right]$$

Usando la desigualdad de Holder se tiene que

$$\mathcal{R}(\mathcal{L}_1, S) \leq \frac{1}{n}\mathbb{E}\left[\sup_{\|w\|_1\leq R_1} \|w\|_1 \cdot \left\|\sum_{i=1}^{n} \sigma_i X_i\right\|_\infty\right] + \frac{1}{n}\mathbb{E}\left[\sup_{|b|\leq r} |b| \left|\sum_{i=1}^{n} \sigma_i\right|\right]$$

El segundo término ya lo calculamos antes, y del primer término nos resulta que

$$\mathcal{R}(\mathcal{L}_1, S) \leq \frac{R_1}{n}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \sigma_i X_i\right\|_\infty\right] + \frac{r}{\sqrt{n}}$$

Pero $\|\sum_{i=1}^n \sigma_i X_i\|_\infty = \max_{j\in\{1,\dots,d\}} |\sum_{i=1}^n \sigma_i X_{ij}|$. Consideremos A un conjunto de vectores ($A_i \in A$ para todo i en $\{1,\dots,d\}$), con $A_i = \{x_{1i},\dots,x_{ni}\}$, esto es, es un conjunto de vectores donde cada vector agrupa una característica sobre todos los datos. Entonces $\|\sum_{i=1}^n \sigma_i X_i\|_\infty = \max_{a\in A} |\sum_{i=1}^n \sigma_i \cdot a_i|$. Pero podemos sacarnos el valor absoluto asumiendo las dos posibilidades de σ_i, esto es, que valga -1 o 1. Así resulta que $\|\sum_{i=1}^n \sigma_i X_i\|_\infty = \max_{a\in A\cup -A} \sum_{i=1}^n \sigma_i \cdot a_i$. Con todo lo anterior cambiamos nuestros datos, y por tanto

$$\mathcal{R}(\mathcal{L}_1, S) \leq R_1 \mathbb{E}\left[\frac{1}{n} \max_{a\in A\cup -A} \sum_{i=1}^n \sigma_i \cdot a_i\right] + \frac{r}{\sqrt{n}}$$

Y podemos ver los a_i como una función constante, y tenemos una cantidad finita de cada una de ellas, y el primer término corresponde a una complejidad de Rademacher, por lo que podemos usar el lema de Massart para concluir que

$$\mathcal{R}(\mathcal{L}_1, S) \leq R_1 \sqrt{\frac{2L_\infty^2 \ln\left(|A\cup -A|\right)}{n}} + \frac{r}{\sqrt{n}}$$

Desarrollando resulta en

$$\mathcal{R}(\mathcal{L}_1, S) \leq L_\infty R_1 \sqrt{\frac{\ln\left(2d\right)}{n}} + \frac{r}{\sqrt{n}}$$

Demostrando lo pedido.

Pregunta 2

(a) Para este caso es un poco más complicado así que no lo haré directamente y usaré la linealidad de la complejidad de Rademacher. Consideremos

$$\mathcal{F} = \left\{ h \in \mathbb{R}^d \to \mathbb{R} \mid h(x) = \sum_{j=1}^m w_j \ell_j(x) : \|w\|_1 \leq R_1, \ell_1, \dots, \ell_m \in \mathcal{L} \right\}$$

$$\mathcal{G} = \left\{ h \in \mathbb{R}^d \to \mathbb{R} \mid h(x) = b : |b| \leq r \right\}$$

Por la propiedad de contracción (para la función σ) y la propiedad de sumatoria de las complejidades de Rademacher, tenemos que

$$\mathcal{R}(\mathcal{H}, S) \leq L\left(\mathcal{R}(\mathcal{F}, S) + \mathcal{R}(\mathcal{G}, S)\right).$$

Sin embargo, como la función es lineal el máximo se debe alcanzar en la frontera, y por tanto considerando

$$\overline{\mathcal{F}} = \left\{ h \in \mathbb{R}^d \to \mathbb{R} \mid h(x) = \sum_{j=1}^m w_j \ell_j(x) : \|w\|_1 = R_1, \ell_1, \dots, \ell_m \in \mathcal{L} \right\}$$

Se tiene que

$$\mathcal{R}(\mathcal{F}, S) = \mathcal{R}(\overline{\mathcal{F}}, S)$$

Pero notando que si sabemos la norma de w podemos descomponer la suma en

$$\sum_{j=1}^{m} w_j \ell_j(x) = \sum_{j=1:w_j \geq 0}^{n} w_j(\ell_j(x) - 0) + \sum_{j=1:w_j \geq 0}^{n} |w_j|(0 - \ell_j(x))$$

Lo cual claramente corresponde a una combinación convexa de $\mathcal{L} - \mathcal{L}$, puesto que por hipótesis la función $0 \in \mathcal{L}$. Por tanto tenemos que

$$\mathcal{R}(\mathcal{F}, S) \leq R_1 \cdot \mathcal{R}(conv(\mathcal{L} - \mathcal{L}), S)$$

Usando la propiedad de la envolvente convexa

$$\mathcal{R}(\mathcal{F}, S) \leq R_1 \cdot \mathcal{R}(\mathcal{L} - \mathcal{L}, S)$$

Usando la propiedad de la suma

$$\mathcal{R}(\mathcal{F}, S) \leq R_1 \cdot \mathcal{R}(\mathcal{L}, S) + \leq R_1 \cdot \mathcal{R}(-\mathcal{L}, S)$$

Y usando la propiedad de multiplicación por un escalar

$$\mathcal{R}(\mathcal{F}, S) \leq 2R_1 \cdot \mathcal{R}(\mathcal{L}, S)$$

Y notando que $\mathcal{R}(\mathcal{G}, S))$ ya lo habíamos calculado en secciones anteriores de la tarea, podemos concluír que

$$\mathcal{R}(\mathcal{H}, S) \leq L \left(2R_1 \cdot \mathcal{R}(\mathcal{L}, S) + \frac{r}{\sqrt{n}} \right).$$

Tal como queríamos demostrar

(b) Para demostrarlo comencemos con el caso base. Digamos que $p = 1$, como la última capa es lineal entonces tiene constante de Lipchitz igual a 1, y por lo demostrado antes se concluye que

$$\mathcal{R}(\mathcal{H}_{NN}^1, S) \leq 2RL\mathcal{R}(\mathcal{H}_{NN}^0, S) + \frac{r}{\sqrt{n}}.$$

Ahora como hipótesis de inducción supongamos que se cumple para $p-1$ con $p \geq 2$, esto es,

$$\mathcal{R}(\mathcal{H}_{NN}^{p-1}, S) \leq 2RL \cdot \mathcal{R}(\mathcal{H}_{NN}^{p-2}, S) + \frac{r}{\sqrt{n}}.$$

Ahora queremos demostrarlo para p. Por definición tenemos que

$$\mathcal{R}(\mathcal{H}_{NN}^{p}, S) = \mathbb{E}\left[\sup_{h\in\mathcal{H}_{NN}^{p}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)\right].$$

Pero recordemos que $h(x) = \sigma^p(W^p h^{p-1}(x) + b^p)$. Y notando que $h^{p-1} \in \mathcal{H}_{NN}^{p-1}$ podemos usar lo mostrado en el punto anterior para concluír que

$$\mathcal{R}(\mathcal{H}_{NN}^{p}, S) \leq 2LR \cdot \mathcal{R}(\mathcal{H}_{NN}^{p-1}, S) + \frac{r}{\sqrt{n}}$$

Y por tanto hemos demostrado la primera parte como inducción, tal como se solicitó. Para demostrar la segunda parte basta usar esto p veces consecutivas. Por tanto, podemos aplicar esta relación recursivamente. Definimos $\mathcal{R}^{(p)} = \mathcal{R}(H_{NN}^{(p)}, S)$. Si aplicamos varias veces, obtenemos

$$\mathcal{R}^{(p)} \leq h \cdot r \cdot \frac{1}{\sqrt{n}} + 2RL \cdot \left(h \cdot r \cdot \frac{1}{\sqrt{n}} + 2RL \cdot \mathcal{R}^{(p-2)}\right)$$

Continuamos aplicando la recursión hasta llegar a $\mathcal{R}^{(0)}$:

$$\mathcal{R}^{(p)} \leq h \cdot r \cdot \frac{1}{\sqrt{n}} \sum_{k=0}^{p} (2RL)^k + (2RL)^p \mathcal{R}^{(0)}$$

Finalmente, de lo demostrado en la 1b sabemos que

$$\mathcal{R}^{(0)} \leq L_\infty R_1 \sqrt{\frac{\ln(2d)}{n}} + \frac{r}{\sqrt{n}}$$

Y reemplazando resulta en

$$R(H_{NN}^{(p)}, S) \leq \frac{1}{\sqrt{n}}\left(r\sum_{k=0}^{p}(2RL)^k + 2R(2RL)^p \max_i \|x_i\|_\infty \sqrt{2\log(2d)}\right)$$

Tal como queríamos demostrar

Pregunta 3

(a) Podemos calcular la función generadora de momentos por definición, tenemos que

$$\mathbb{E}[\exp(\lambda x)] = \int_a^b \exp(\lambda x) \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\int_a^b \exp(\lambda x)\, dx$$

Resolviendo la integral, se obtiene

$$\mathbb{E}[\exp(\lambda x)] = \frac{\exp(\lambda b) - \exp(\lambda a)}{\lambda(b-a)}, \quad \lambda \neq 0$$

Para $\lambda = 0$, tenemos que $\mathbb{E}[\exp(0 \cdot x)] = 1$

(b) Para probar que (y_k) es una martingala con respecto a (y_k) lo haremos verificando las propiedades elementales.

1. y_k es $\sigma(x_1, \ldots, x_k)$-medible: Notemos que y_k se construye en base a $x_1, \ldots, x_k$ solo usando multiplicación y resta (ambas funciones trivialmente medibles). Además, es claro que x_k es $\sigma(x_1, \ldots, x_k)$-medible y por tanto al ser y_k una composición de funciones $\sigma(x_1, \ldots, x_k)$-medibles, resulta ser $\sigma(x_1, \ldots, x_k)$-medible.

2. $\mathbb{E}[|y_k|] < +\infty$: Dado que $x_k \in [0, 1]$, es claro que $1 - x_k \in [0, 1]$, por tanto, $y_k = 2^k(1 - x_k) \leq 2^k$. Esto implica que $\mathbb{E}[|Y_k|] \leq 2^k$, lo cual es finito para cada k finito.

3. $\mathbb{E}[y_{k+1}|x_1, \ldots, x_k] = y_k$: Por definición tenemos que

   $$y_{k+1} = 2^{k+1}(1 - x_{k+1})$$

   Para calcular $\mathbb{E}[y_{k+1} \mid x_1, \ldots, x_k]$, observamos que, dado x_k, la variable x_{k+1} sigue una distribución uniforme en el intervalo $[x_k, 1]$. Por lo tanto

   $$\mathbb{E}[x_{k+1} \mid x_k] = \frac{x_k + 1}{2}$$

   Por lo que

   $$\mathbb{E}[y_{k+1} \mid x_k] = \mathbb{E}[2^{k+1}(1 - x_{k+1}) \mid x_k] = 2^{k+1}\left(1 - \frac{x_k + 1}{2}\right)$$

   Simplificando resulta que

   $$\mathbb{E}[y_{k+1} \mid x_k] = 2^{k+1} \cdot \frac{1 - x_k}{2} = 2^k(1 - x_k) = y_k$$

   Tal como queríamos demostrar

Como verificamos la propiedad que lo caracterizan como una martingala, concluímos que efectivamente corresponde a una martingala.

(c) Notemos que la diferencia $d_{k+1} = y_{k+1} - y_k$ corresponde, por definición, a

$$d_{k+1} = 2^{k+1}(1 - d_{k+1}) - 2^k(1 - x_k)$$

Factorizamos

$$d_{k+1} = 2^k\left(2(1 - x_{k+1}) - (1 - x_k)\right)$$

Desarrollando

$$d_{k+1} = 2^k\left(2 - 2x_{k+1} - 1 + x_k\right) = 2^k(1 - 2x_{k+1} + x_k)$$

Ahora calculemos lo pedido, esto es

$$\mathbb{E}\left[\exp(\lambda d_{k+1}) \mid x_1, \ldots, x_k\right]$$

Condicionado a x_k, sabemos que $x_{k+1} \sim \text{Unif}([x_k, 1])$. Entonces

$$\mathbb{E}\left[\exp(\lambda d_{k+1}) \mid x_k\right] = \mathbb{E}\left[\exp\left(\lambda \cdot 2^k(1 - 2x_{k+1} + x_k)\right) \mid x_k\right]$$

Sacamos las constantes (como está condicionado a x_k entonces dicho valor ya no es aleatorio)

$$\mathbb{E}\left[\exp(\lambda d_{k+1}) \mid x_k\right] = \exp\left(\lambda \cdot 2^k(1 + x_k)\right) \cdot \mathbb{E}\left[\exp(-2\lambda \cdot 2^k x_{k+1}) \mid x_k\right]$$

Usemos que $x_{k+1} \mid x_k \sim \text{Unif}([x_k, 1])$, entonces

$$\mathbb{E}\left[\exp(-2\lambda \cdot 2^k x_{k+1}) \mid x_k\right] = \int_{x_k}^{1} \exp(-2\lambda \cdot 2^k x) \cdot \frac{1}{1 - x_k}\, dx$$

Para calcular la integral notamos que la primitiva de $\exp(-2\lambda \cdot 2^k x)$ es $-\frac{1}{2\lambda \cdot 2^k}\exp(-2\lambda \cdot 2^k x)$, de lo que resulta

$$\int_{x_k}^{1} \exp(-2\lambda \cdot 2^k x) \cdot \frac{1}{1 - x_k}\, dx = \frac{-1}{2\lambda \cdot 2^k(1 - x_k)}\left(\exp(-2\lambda \cdot 2^k) - \exp(-2\lambda \cdot 2^k x_k)\right)$$

Sustituyendo todo

$$\mathbb{E}\left[\exp(\lambda d_{k+1}) \mid x_k\right] = \frac{\exp\left(\lambda \cdot 2^k(1 + x_k)\right)}{\lambda 2^{k+1}(1 - x_k)}\left(\exp(-2\lambda \cdot 2^k x_k) - \exp(-2\lambda \cdot 2^k)\right)$$

Encontrando la función generadora de momentos solicitada.