Question 1

Ben Eliav, Eyal Tadmor

March 22, 2024

1.1

For any hypothesis class of size 1, the Rademacher complexity for l is $E_\sigma\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$ (removing the maximization). Using the linearity of the expectation, we get that the Rademacher complexity is $E_\sigma\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right] = \frac{1}{m}\sum_{i=1}^{m}E_\sigma\left[\sigma_i\right]f(z_i) = \frac{1}{m}\sum_{i=1}^{m}0 = 0$. This is the minimum possible value because for any hypothesis class $\mathcal{H}$, with $\mathcal{H}' = \{h\} \subseteq \mathcal{H}$, defining $F = l \circ \mathcal{H}, F' = l \circ \mathcal{H}'$, we get the following:

$$\forall \sigma_1, \ldots, \sigma_m : \sup_{f\in F}\sum_{i=1}^{m}\sigma_i f(z_i) \geq \sup_{f\in F'}\sum_{i=1}^{m}\sigma_i f(z_i)$$

$$\implies E_\sigma\left[\frac{1}{m}\sup_{f\in F}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \geq E_\sigma\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right] = 0$$

where the first inequality is due to the fact that the supremum of a set is greater than the supremum of any of its subsets. Therefore, the minimum Rademacher complexity is 0.
For a hypothesis class $\mathcal{H}$ of size 2^m that contains all possible hypotheses, we can get the maximum Rademacher complexity. We will set some sequence $\sigma_1, \ldots \sigma_m$. Because $\mathcal{H}$ contains all possible hypotheses, we can find a hypothesis $h \in \mathcal{H}$ such that $f(z_i) = \sigma_i$. Therefore, the Rademacher complexity is $E_\sigma\left[\frac{1}{m}\sup_{f\in F}\sum_{i=1}^{m}\sigma_i f(z_i)\right] = E_\sigma\left[\frac{1}{m}m\right] = 1$. This is the maximal possible value of the Rademacher complexity for a $\{-1, +1\}$ loss l because each member of the sum is always equal to or less than 1 (product of two numbers between -1 and 1).

Therefore, in the case of $\{-1, +1\}$ loss, the minimum Rademacher complexity is 0 (when there is only one hypothesis) and the maximum Rademacher complexity is 1 (when looking at all possible hypotheses). Notice that we usually do not want to use loss functions that can return negative numbers, but if we did then we know how to bound the Rademacher complexity.

1.2

Note that the $\{0,1\}$ loss $l(z_i, h)$ can be viewed as $\frac{1-h(x_i)y_i}{2}$. Plugging it into the Rademacher complexity definition, we can get:

$$
\begin{aligned}
R(L \circ S) &= E_\sigma \left[\frac{1}{m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i l(z_i, h) \right] \\
&= E_\sigma \left[\frac{1}{m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i \frac{1 - h(x_i)y_i}{2} \right] \\
&= E_\sigma \left[\frac{1}{2m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i - \sigma_i h(x_i)y_i \right] \\
&\overset{(*)}{=} \frac{1}{2m} \sum_{i=1}^{m} E_\sigma \left[\sigma_i\right] + E_\sigma \left[\frac{1}{2m} \sup_{h \in H} \sum_{i=1}^{m} -\sigma_i h(x_i)y_i \right] \\
&= \frac{1}{2m} \sum_{i=1}^{m} 0 + \frac{1}{2} E_\sigma \left[\frac{1}{m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i)y_i \right] \\
&= \frac{1}{2} E_\sigma \left[\frac{1}{m} \sup_{h \in H} \sum_{i:\, y_i = 1} \sigma_i h(x_i) + \sum_{j:\, y_j \neq 1} -\sigma_j h(x_j) \right] \\
&= \frac{1}{2} \sum_{\sigma} P(\sigma) \left[\frac{1}{m} \sup_{h \in H} \sum_{i:\, y_i = 1} \sigma_i h(x_i) + \sum_{j:\, y_j \neq 1} -\sigma_j h(x_j) \right] \\
&\overset{(**)}{=} \frac{1}{2} \sum_{\sigma^-} P(\sigma^-) \left[\frac{1}{m} \sup_{h \in H} \sum_{i:\, y_i = 1} \sigma_i^- h(x_i) + \sum_{j:\, y_j \neq 1} \sigma_j^- h(x_j) \right] \\
&= \frac{1}{2} \sum_{\sigma} P(\sigma) \left[\frac{1}{m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] \\
&= \frac{1}{2} E_\sigma \left[\frac{1}{m} \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \frac{1}{2} R(H \circ S)
\end{aligned}
$$

where $(*)$ is due to the fact that the sum on σ_i does not depend on h and thus can be extracted from the supremum.
To explain $(**)$, we define σ as a sequence of m Rademacher random variables. We can define σ^- as a transformation on some σ such that $\sigma_i^- = -\sigma_i$ for all i such that $y_i = -1$. Because σ_i are i.i.d, the probability of $\sigma_i = 1$ is the same as the probability of $\sigma_i^- = -1$. Therefore, $P(\sigma) = P(\sigma^-)$. Also note that the transformation $\sigma \to \sigma^-$ is a bijection, so summing over all σ is the same as summing over all σ^-.