**MIT 9.520/6.860: Statistical Learning Theory and Applications** **Fall 2022**

Problem Set 2

*Student Name: Aniruddha Deshpande*

# Problem 1

## Problem 1.1

I will proceed with this problem in the following way. First, I will derive the analytical form for the minimizing function $f_\rho = \arg\min_f \mathcal{E}(f)$. I will do this by using the same approach as used in Q1 of pset 1 (or section 1.5.1 of the book) where , to find the $f$ that minimizes $L(f)$ as given below (and $L_x(a) = l(y,a)dP(y|x)$) it is sufficient to minimize $L_x(f(x))$ over $f$ and that will give us the same optimal function.

$$L(f) = \int dP_X(x) L_x(f(x)) \tag{2.1}$$

Then, once I have this, I will characterize $|R(f) - R(f_\rho)|$. To find $f_\rho$ -

$$\begin{aligned}\mathcal{E}(f) &= \int_{X\times Y} (f(x)-y)^2 p(x,y)\,dx\,dy \\ &= \int_X [(f(x)-1)^2 p(x,1) + (f(x)+1)^2 p(x,-1)]\,dx\end{aligned}$$

As mentioned above, $f_\rho$ can be found by just minimizing the term inside the integral. This can be done by setting the derivative of this with respect to $f$ to 0.

$$2(f_\rho(x)-1)p(x,1) + 2(f_\rho(x)+1)p(x,-1) = 0$$

$$f_\rho(x)p(x) = p(x,1) - p(x,-1)$$

$$f_\rho(x) = \frac{p(x,1) - p(x,-1)}{p(x)}$$

Note, that the optimal function that minimizes the expected risk is a pretty intuitive function, which gives a positive value (and hence predicts $\hat{y} = +1$) when $p(x,1) > p(x,-1)$ and a negative value when $p(x,1) < p(x,-1)$.

Now, to characterize the quantity $|R(f) - R(f_\rho)|$. For any arbitrary function $f \in \mathcal{F}$ and $f_\rho$ we can divide the input space $\mathcal{X}$ into 4 regions $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4$ such that :-

$$
\begin{aligned}
\mathcal{X}_1 &= \{x \in \mathcal{X} | f_\rho(x) \geq 0, f(x) \geq 0\} \\
\mathcal{X}_2 &= \{x \in \mathcal{X} | f_\rho(x) < 0, f(x) \geq 0\} \\
\mathcal{X}_3 &= \{x \in \mathcal{X} | f_\rho(x) \geq 0, f(x) < 0\} \\
\mathcal{X}_4 &= \{x \in \mathcal{X} | f_\rho(x) < 0, f(x) < 0\} \\
\mathcal{X} &= \mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3 \cup \mathcal{X}_4
\end{aligned}
$$

Using this, we can write the $R(f)$ and $R(f_\rho)$ as :-

$$
\begin{aligned}
R(f) &= \mathbb{P}(sign(f(x)) \neq y) \\
&= \mathbb{P}(\{f(x) \geq 0, y = -1\} \cup \{f(x) < 0, y = +1\}) \\
&= \mathbb{P}(f(x) \geq 0, y = -1) + \mathbb{P}(f(x) < 0, y = +1) \\
&= \int_{\mathcal{X}_1 \cup \mathcal{X}_2} p(x, -1)\, dx + \int_{\mathcal{X}_3 \cup \mathcal{X}_4} p(x, 1)\, dx
\end{aligned}
$$

Similarly, for $R(f_\rho)$ , we have:-

$$
\begin{aligned}
R(f_\rho) &= \mathbb{P}(sign(f_\rho(x)) \neq y) \\
&= \mathbb{P}(\{f_\rho(x) \geq 0, y = -1\} \cup \{f_\rho(x) < 0, y = +1\}) \\
&= \mathbb{P}(f_\rho(x) \geq 0, y = -1) + \mathbb{P}(f_\rho(x) < 0, y = +1) \\
&= \int_{\mathcal{X}_1 \cup \mathcal{X}_3} p(x, -1)\, dx + \int_{\mathcal{X}_2 \cup \mathcal{X}_4} p(x, 1)\, dx
\end{aligned}
$$

Using the above we get:-

$$
\begin{aligned}
R(f) - R(f_\rho) &= \int_{\mathcal{X}_2} p(x,-1)\,dx - \int_{\mathcal{X}_2} p(x,1)\,dx + \int_{\mathcal{X}_3} p(x,1)\,dx - \int_{\mathcal{X}_3} p(x,-1)\,dx \\
&= \int_{\mathcal{X}_3} [p(x,1) - p(x,-1)]\,dx + \int_{\mathcal{X}_2} [p(x,-1) - p(x,1)]\,dx \\
&= \int_{\mathcal{X}_3} |p(x,1) - p(x,-1)|\,dx + \int_{\mathcal{X}_2} |p(x,1) - p(x,-1)|\,dx \\
&= \int_{\mathcal{X}_2 \cup \mathcal{X}_3} |p(x,1) - p(x,-1)|\,dx \\
&= \int_{\mathcal{X}_f} |p(x,1) - p(x,-1)|\,dx \\
&= \int_{\mathcal{X}_f} \frac{|p(x,1) - p(x,-1)|}{p(x)} p(x)\,dx \\
&= \int_{\mathcal{X}_f} \frac{|p(x,1) - p(x,-1)|}{p(x)}\,d\rho(x) \\
&= \int_{\mathcal{X}_f} |f_\rho(x)|\,d\rho(x)
\end{aligned}
$$

Where, the third equality is because $\forall x \in \mathcal{X}_3$ we have $f_\rho \geq 0$, which implies $p(x,1) \geq p(x,-1)$. Similarly $\forall x \in \mathcal{X}_3$ we have $f_\rho < 0$, which implies $p(x,1) < p(x,-1)$. The last equality follows from the definitions of $\mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_f$.

## Problem 1.2

Now, to prove the inequalities. The first inequality is straightforward from the observation that for the space $\mathcal{X}_f$ the functions $f$ and $f_\rho$ have the opposite signs. This means that $\forall x \in \mathcal{X}_f$ the following is always true:

$$
\begin{aligned}
|f_\rho(x)| &\leq |f_\rho(x) - f(x)| \\
|f_\rho(x)|p(x) &\leq |f_\rho(x) - f(x)|p(x)
\end{aligned}
$$

Where, the second inequality holds because $p(x)$ is a positive quantity.

Integrating this inequality on both sides over $\mathcal{X}_f$ will preserve the inequality and hence we have.

$$
\int_{\mathcal{X}_f} |f_\rho(x)|p(x)\,dx \leq \int_{\mathcal{X}_f} |f_\rho(x) - f(x)|p(x)\,dx
$$

To prove the second inequality, I will use the Jensen's inequality for concave functions (specifically the square root function).

$$\begin{aligned}\int_{\mathcal{X}_f} |f_\rho(x) - f(x)| p(x)\, dx &= \int_{\mathcal{X}_f} \sqrt{|f_\rho(x) - f(x)|^2} p(x)\, dx \\ &= \mathbb{E}_{p(x)}[\sqrt{|f_\rho(x) - f(x)|^2}] \\ &\leq \sqrt{\mathbb{E}_{p(x)}(|f_\rho(x) - f(x)|^2)}\end{aligned}$$

Where as mentioned above, the last inequality is using the Jensen's inequality for the concave square root function.

## Problem 1.3

This part can be proved by writing out the definition of $\mathcal{E}(f), \mathcal{E}(f)$ and $f_\rho$ as derived in 1.1

$$\begin{aligned}\mathcal{E}(f) - \mathcal{E}(f_\rho) &= \int_{X\times Y} [(f(x)-1)^2 - (f_\rho(x)-1)^2] p(x,y)\, dx\, dy \\ &= \int_X ([(f(x)-1)^2 - (f_\rho(x)-1)^2] p(x,1) + [(f(x)+1)^2 - (f_\rho(x)+1)^2] p(x,-1))\, dx\, dy \\ &= \int_X ([f^2(x) + 1 - 2f(x) - f_\rho^2(x) - 1 + 2f_\rho(x)] p(x,1) \\ &+ [f^2(x) + 1 + 2f(x) - f_\rho^2(x) - 1 - 2f_\rho(x)] p(x,-1))\, dx\, dy \\ &= \int_X [f^2(x)p(x) - f_\rho^2(x)p(x) - 2f(x)(p(x,1) - p(x,-1)) + 2f_\rho(x)(p(x,1) - p(x,-1))]\, dx\, dy \\ &= \int_X [f^2(x)p(x) - f_\rho^2(x)p(x) - 2f(x)(f_\rho(x)p(x)) + 2f_\rho(x)(f_\rho(x)p(x))]\, dx\, dy \\ &= \int_X [f^2(x)p(x) + f_\rho^2(x)p(x) - 2f(x)(f_\rho(x)p(x))]\, dx\, dy \\ &= \int_X [f(x) - f_\rho(x)]^2 p(x)\, dx\, dy \\ &= \mathbb{E}(|f(x) - f_\rho(x)|^2)\end{aligned}$$

Where I get the 5th equality by replacing $(p(x,1) - p(x,-1))$ with $f_\rho(x)p(x)$

# Problem 2

First, combining the two inequalities given in the question into 1, we get:

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}(U)| > \epsilon) \leq 2\exp\{-\frac{2n\epsilon^2}{(a-b)^2}\}$$

We also have $\forall j \in [n]$ :-

$$\mathbb{L}(f_j) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(f_j(x), y)]$$
$$\hat{\mathbb{L}}(f_j) = \frac{1}{n}\sum_{i=1}^{n} l(f_j(x_i), y_i)$$

Note, that for any chosen function, $\hat{\mathbb{L}}$ is a sum of iid random variables, given that $\{x_i, y_i\}$ are sampled iid.

Now, I will try to bound the term given in the question. I will use the following facts. First, I will use the fact that, given 2 events A and B; $\mathbb{P}(A) \leq \mathbb{P}(A \cup B)$ . I will then use the union bound, followed by Hoeffding's inequality mentioned above.

$$\begin{aligned}
\mathbb{P}(\max_{f\in\mathcal{F}} |\mathbb{L}(f) - \hat{\mathbb{L}}(f)| > \epsilon) &\leq \mathbb{P}(\bigcup_{i=1}^{n}\{|\mathbb{L}(f_i) - \hat{\mathbb{L}}(f_i)| > \epsilon\}) \\
&\leq \sum_{i=1}^{n} \mathbb{P}(|\mathbb{L}(f_i) - \hat{\mathbb{L}}(f_i)| > \epsilon) \\
&\leq 2N \exp\{-\frac{2n\epsilon^2}{(C+M)^4}\}
\end{aligned}$$

The last inequality comes from the Hoeffding's bound. The denominator in the exponential $(a-b)^2$ is essentially the square of the range of the random variable $U_i$. In my application of this bound, $U_i \to l(f_j(x_i), y_i) = (f_j(x_i) - y_i)^2$. The minimum value of $l(.,.)$ is 0. The maximum value of this loss would be achieved when $y = M$ and $f_j(x_i) = -C$. So, the range of $l(f_j(x_i), y_i)$ is $[0, (C+M)^2]$. Substituting this range into the Hoeffding bound we get the result.

## Problem 2.2

To solve this question, I will first use the following idea.

$$\mathbb{P}(\max_{f\in\mathcal{F}} |\mathbb{L}(f) - \hat{\mathbb{L}}(f)| \geq \epsilon) \geq \mathbb{P}(|\mathbb{L}(f) - \hat{\mathbb{L}}(f)| \geq \epsilon) \qquad \forall f \in \mathcal{F}$$

This is true, because the event that the max value of some function is greater than a given value, is included within the event that the value of the function at any other point is greater than a given value, ie. if the event on the right hand side takes place, then the event on the left hand side taking place is implied. Using this, we have

$$
\begin{aligned}
\mathbb{P}(|\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n)| \geq \epsilon) &\leq \mathbb{P}(\max_{f \in \mathcal{F}} |\mathbb{L}(f) - \hat{\mathbb{L}}(f)| \geq \epsilon) \\
&\leq 2N \exp\{-\frac{2n\epsilon^2}{(C+M)^4}\}
\end{aligned}
$$

Splitting the LHS, we get

$$
\mathbb{P}(\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) \geq \epsilon) + \mathbb{P}(\hat{\mathbb{L}}(\hat{f}_n) - \mathbb{L}(\hat{f}_n) \geq \epsilon) \leq 2N \exp\{-\frac{2n\epsilon^2}{(C+M)^4}\}
$$

The second term $\mathbb{P}(\hat{\mathbb{L}}(\hat{f}_n) - \mathbb{L}(\hat{f}_n) = 0$ , as $\hat{\mathbb{L}}(\hat{f}_n) - \mathbb{L}(\hat{f}_n) \leq 0$. This is because ... Now, using this we can see that:

$$
\begin{aligned}
1 - \mathbb{P}(\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) \geq \epsilon) = \mathbb{P}(\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) < \epsilon) \\
\geq 1 - 2N \exp\{-\frac{2n\epsilon^2}{(C+M)^4}\}
\end{aligned}
$$

Now, given the framing of the question, we can see that $\delta \to 2n \exp\{-\frac{2n\epsilon^2}{(C+M)^4}\}$. So, the problem becomes finding $\epsilon$ in terms of $\delta$ which can be given by:-

$$
\epsilon(n, M, \gamma) = \sqrt{\frac{(C+M)^4}{2n} \log(\frac{2N}{\delta})}
$$

## Problem 2.3

Starting with the LHS in the question and adding/subtracting a few terms we get:-

$$
\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) + \hat{\mathbb{L}}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) + \hat{\mathbb{L}}(f_{\mathcal{F}}) - \hat{\mathbb{L}}(f_{\mathcal{F}}) < \mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) - \mathbb{L}(\hat{f}_{\mathcal{F}}) + \hat{\mathbb{L}}(f_{\mathcal{F}})
$$

The inequality is because $\hat{\mathbb{L}}(\hat{f}_n) - \hat{\mathbb{L}}(f_{\mathcal{F}}) < 0$ as $\hat{f}_n$ is the minimizer of the empirical risk. Now I will proceed as follows:-

$$\begin{aligned}
\mathbb{P}(\{\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) > \epsilon\} \cup \{\hat{\mathbb{L}}(f_{\mathcal{F}}) - \mathbb{L}(\hat{f}_{\mathcal{F}}) > \epsilon\}) &\leq \mathbb{P}(\max_{f \in \mathcal{F}} |\mathbb{L}(f) - \hat{\mathbb{L}}(f)| > \epsilon) \\
&\leq \delta \\
1 - \mathbb{P}(\{\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) > \epsilon\} \cup \{\hat{\mathbb{L}}(f_{\mathcal{F}}) - \mathbb{L}(\hat{f}_{\mathcal{F}}) > \epsilon\}) &\geq 1 - \delta \\
\mathbb{P}(\{\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) > \epsilon\}^C \cap \{\hat{\mathbb{L}}(f_{\mathcal{F}}) - \mathbb{L}(\hat{f}_{\mathcal{F}}) > \epsilon\}^C) &\geq 1 - \delta \\
\mathbb{P}(\{\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) \leq \epsilon\} \cap \{\hat{\mathbb{L}}(f_{\mathcal{F}}) - \mathbb{L}(\hat{f}_{\mathcal{F}}) \leq \epsilon\}) &\geq 1 - \delta \\
\mathbb{P}(\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) - \mathbb{L}(\hat{f}_{\mathcal{F}}) + \hat{\mathbb{L}}(f_{\mathcal{F}}) \leq 2\epsilon) &\geq \mathbb{P}(\{\mathbb{L}(\hat{f}_n) - \hat{\mathbb{L}}(\hat{f}_n) \leq \epsilon\} \cap \{\hat{\mathbb{L}}(f_{\mathcal{F}}) - \mathbb{L}(\hat{f}_{\mathcal{F}}) \\
&\geq 1 - \delta
\end{aligned}$$

The first inequality is true because of a similar argument as used in 2.2 ; the event that the max of a function within a class is greater than a given value is more likely than any other function within the class being greater than a given value, and the same applies for any 2 functions being greater than a given value. The third line is true using $(A \cup B)^C = A^C \cap B^C$. The last line is true because $\mathbb{P}(a + b < 2c) \geq \mathbb{P}(\{a < c\} \cap \{b < c\})$.

# Problem 3

## Problem 3.1

This part can be proved using the triangle inequality as mentioned in the hint. I will start with the definition of replace-one stability and proceed from there:

$$\begin{aligned}
|V(f_{\mathcal{S}}, z') - V(f_{\mathcal{S}_{i,z}}, z')| &= |V(f_{\mathcal{S}}, z') - V(f_{\mathcal{S}_i}, z') + V(f_{\mathcal{S}_i}, z') - V(f_{\mathcal{S}_{i,z}}, z')| \\
&\leq |V(f_{\mathcal{S}}, z') - V(f_{\mathcal{S}_i}, z')| + |V(f_{\mathcal{S}_i}, z') - V(f_{\mathcal{S}_{i,z}}, z')| \\
&\leq \beta + \beta \\
&= 2\beta
\end{aligned}$$

where, the first inequality is using the triangle inequality and the second inequality is using the fact that the given algorithm is $\beta$-uniform leave one out stable.

## Problem 3.2

Given that $\mathcal{Z}$ has finite support. Let $|\mathcal{Z}| = N > n$ . Now given that the learning algorithm is $\lambda$ L1-stable , then we have $\forall z \in \mathcal{Z}$ and any distributions $p, q$ :-

$$|V(f_p, z') - V(f_q, z')| \leq \lambda ||p - q||_1$$

Now, to prove that the above implies uniform leave one-out stability it would be sufficient to show that $V(f_{\mathcal{S}}, z') = V(f_p, z')$ and $V(f_{\mathcal{S}_i}, z') = V(f_q, z')$ for some choice of $p$ and $q$. Note that any dataset $\mathcal{S}$ can be fully identified by a set of indices $\mathcal{S}_{ind} \subset [N], |\mathcal{S}_{ind}| = n$ It can be observed that giving the data set $\mathcal{S}$ to the algorithm is equivalent to feeding the algorithm with data $z \in \mathcal{Z}$ such that $p_j = \frac{1}{n}$ , $\forall j \in S_{ind}$ and 0 otherwise. Similarly giving the data set $\mathcal{S}_i$ to the algorithm is equivalent to feeding the algorithm with data $z \in \mathcal{Z}$ such that $q_j = \frac{1}{n-1}$ , $\forall j \in S_i^{ind}$ and 0 otherwise. Also noted that given the definitions of the sets $S^{ind}$ and $S_i^{ind}$ , $S_i^{ind} \subset S^{ind}$ , there is only one index in the former set which is not included in the latter. So, using this we can get the upper bound on the p,q stability as :

$$\begin{aligned}\lambda||p-q||_1 &= \lambda[(n-1)(\frac{1}{n-1} - \frac{1}{n}) + \frac{1}{n}] \\ &= \frac{2\lambda}{n}\end{aligned}$$

## Problem 3.3

I couldn't figure out a good enough solution for this problem. We tried using an approach where we defined a switch $\epsilon$ stability where instead of fully swapping out a sample from the data set we slightly re-weight the probabilities by adding and subtracting $\epsilon$ from the probabilities of 2 samples. But this didn't work out. Curious to see the solution to this!

## Problem 3.4

First, note that the LHS of the inequality to be proven follows:

$$\begin{aligned}\mathbb{E}_{\mathcal{S},z}[|V(f_{\mathcal{S}}, z') - V(f_{\mathcal{S}_i}, z')|] &= \mathbb{E}_{\mathcal{S},z}(\mathbb{1}_{\{f_{\mathcal{S}_i(x)} \neq f_{\mathcal{S}_i}(x)\}}) \\ &= \mathbb{P}(f_{\mathcal{S}}(x) \neq f_{\mathcal{S}_i(x)})\end{aligned}$$

The first equality is true because of the definition of $V(f, z)$. The term will be zero every time $f_{\mathcal{S}}$ and $f_{\mathcal{S}_i}$ is the same give the same prediction and 1 otherwise and hence the term can be replaced by the appropriate indicator function.

Then, the proof follows from the observation that the prediction of $f_{\mathcal{S}}$ and $f_{\mathcal{S}_i}$ can differ only if the removed sample $z_i$ is one of the k-nearest neighbours of $z$. By the symmetry of the algorithm I can argue that the probability of this event is $\frac{k}{n}$ . The inequality in the result given is because, even when the chosen sample to be removed from the data set is one of the k-nearest neighbours the output of the algorithm doesn't not necessarily change, so the probability of the outputs of $f_{\mathcal{S}}$ and $f_{\mathcal{S}_i}$ being different is at most $\frac{k}{n}$.

# Problem 4

## Problem 4.1

I will proceed with derivation for the first bound as in lecture 6.

$$
\begin{aligned}
\mathcal{R}_X(\mathcal{F}_W) &= \mathbb{E}_\sigma\Big(\sup_{||w||\leq W}\big(\frac{1}{n}\sum_i \sigma_i < w, x_i >\big)\Big) \\
&= \mathbb{E}_\sigma\Big(\sup_{||w||\leq W}\big(< w, \frac{1}{n}\sum_i \sigma_i x_i >\big)\Big) \\
&\leq \mathbb{E}_\sigma\Big(\sup_{||w||\leq W} ||w||.||\frac{1}{n}\sum_i \sigma_i x_i||\Big) \\
&\leq W\mathbb{E}_\sigma\Big(\sqrt{||\frac{1}{n}\sum_i \sigma_i x_i||^2}\Big) \\
&\leq W\sqrt{\mathbb{E}_\sigma\big(||\frac{1}{n}\sum_i \sigma_i x_i||^2\big)} \\
&= W\sqrt{\frac{1}{n^2}\mathbb{E}_\sigma\big(\sum_{i,j} \sigma_i \sigma_i < x_i, x_j >\big)} \\
&= W\sqrt{\frac{1}{n^2}\mathbb{E}_\sigma\big(\sum_i \sigma_i^2 ||x_i^2|| + \sum_{i\neq j} \sigma_i \sigma_i < x_i, x_j >\big)} \\
&= W\sqrt{\frac{1}{n^2}\mathbb{E}_\sigma\big(\sum_i ||x_i^2||\big)} \\
&\leq W\sqrt{\frac{1}{n^2} n B^2} \\
&= \frac{BW}{\sqrt{n}}
\end{aligned}
$$

where, we have the second equality by the linearity of the inner product operator. The first inequality is using the Cauchy-Schwarz inequality. The second inequality comes from the fact that $||w||$ is bounded and the third inequality is the Jensen's inequality applied for the concave square root function. The last inequality comes $||x||$ being bounded.

## Problem 4.2

Using the definitions given in the question, we have:

$$\begin{aligned}
\mathcal{R}_X(\mathcal{F}_1+\mathcal{F}_2) &= \mathbb{E}_\sigma\Big(\sup_{g\in\mathcal{F}_1+\mathcal{F}_2}\big(\frac{1}{n}\sum_i \sigma_i g(x_i)\big)\Big)\\
&= \mathbb{E}_\sigma\Big(\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2}\big(\frac{1}{n}(\sum_i \sigma_i f_1(x_i) + \sum_i \sigma_i f_2(x_i))\big)\Big)\\
&= \mathbb{E}_\sigma\Big(\sup_{f_1\in\mathcal{F}_1}\frac{1}{n}\sum_i \sigma_i f_1(x_i) + \sup_{f_2\in\mathcal{F}_2}\frac{1}{n}\sum_i \sigma_i f_2(x_i)\Big)\\
&= \mathbb{E}_\sigma\Big(\sup_{f_1\in\mathcal{F}_1}\frac{1}{n}\sum_i \sigma_i f_1(x_i)\Big) + \mathbb{E}_\sigma\Big(\sup_{f_2\in\mathcal{F}_2}\frac{1}{n}\sum_i \sigma_i f_2(x_i)\Big))\\
&= \mathcal{R}_X(\mathcal{F}_1) + \mathcal{R}(\mathcal{F}_2)
\end{aligned}$$

where, the second equality is by the structure of the function class $\mathcal{F}_1 + \mathcal{F}_2$. The third equality follow from the linearity of the supremum operator and the fact that the first term only depends on $f_1$ and the second only on $f_2$. The next equality follows from linearity of expectation.