MIT 9.520/6.860: Statistical Learning Theory and Applications **Fall 2018**

Problem Set 2

Student Name: Sebastiani Aguirre-Navarro

Problem 1 1.1

Using the symmetrization lemma

$$\mathbb{E} \max_{f \in \mathcal{F}} \left[\mathbb{E}\, f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \leq 2\, \mathbb{E}\, \hat{R}_n(\mathcal{F}|_{x1:n})$$

In this case $\mathcal{F}|_{x1:n}$ is finite, since we are conditioning on n x-vectors. Then

$$2\, \mathbb{E}\, \hat{R}_n(\mathcal{F}|_{x1:n}) = 2\, \mathbb{E} \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$$

Using Massart's lemma, $\mathbb{E}\left[\max_{f\in\mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i)\right] \leq r\sqrt{2\log|\mathcal{F}|}$ where $r = \max_{f\in\mathcal{F}}||f||_2$, but $\mathcal{F}$ is a class of indicator functions which are binary vectors so $r = \sqrt{n}$. Then,

$$\frac{2}{n}\sqrt{n}\sqrt{2\log|\mathcal{F}|} = c\sqrt{\frac{\log|\mathcal{F}|_{x1:n}|}{n}}$$

The cardinality of $\mathcal{F}|_{x1:n}$ is n vectors, so

$$\mathbb{E} \max_{f \in \mathcal{F}} \left[\mathbb{E} - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \leq c\sqrt{\frac{\log n}{n}}$$

with $c = 2\sqrt{2}$

1.2

With $l(f(x), y) = \mathbb{I}\{f(x) \neq y\} = \frac{1-f(x_i)y_i}{2}$, then,

$$2\, \mathbb{E} \max_{g \in l \circ \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \frac{(1 - f(x_i)y_i)}{2}$$

But $(1-f(x_i)y_i)$ is binary valued between 0 and 2, therefore the same result can be obtained $\sqrt{2}\sqrt{\frac{\log n}{n}}$ Hence the difference is a 2 between the two Rademacher averages.

1.3 With $\hat{f}_n$ as the solution of the ERM, we can study the bounds of $\mathbb{E}\, L(\hat{f}_n) - \hat{L}(f_{\mathcal{F}})$ whose Rademacher averages we calculated. Because we are using the solution of the ERM and the Bayes optimal function $f_{\mathcal{F}}$ then ϵ will be greater than the bound only when:

$$\begin{aligned} c\sqrt{\frac{\log n}{n}} &\leq \epsilon \\ \sqrt{\frac{\log n}{n}} &\leq \epsilon \\ \frac{\log n}{n} &\leq \epsilon^2 \end{aligned} \tag{2.1}$$

Ignoring log factors, $n \geq \mathcal{O}(\epsilon^{-2})$

1.4

Let $p \geq 1$, $B_p^n = x \in \hat{\mathbb{R}}^n : ||x||_p \leq 1$ Using the definition of the dual norm, we have

$$\begin{aligned} n\hat{\mathbb{R}}(B_p^n) &= \mathbb{E} \max_{||x|| \leq_p 1} \langle \epsilon, x \rangle \\ &= \mathbb{E}||\epsilon||_p \\ &= \mathbb{E}(||\epsilon||_p^p)^{\frac{1}{p}} \\ &\leq (\mathbb{E}||\epsilon||_p^p)^{\frac{1}{p}} \end{aligned} \tag{2.2}$$

Because ϵ takes values ± 1, then $\mathbb{E}(||\epsilon||_p^p)^{\frac{1}{p}} \approx n^{\frac{1}{p}}$ then

$$\begin{aligned} \hat{\mathbb{R}}(B_p^n) &\approx \frac{n^{\frac{1}{p}}}{n} \\ &= \mathcal{O}(n^{-\frac{1}{p}}) \end{aligned} \tag{2.3}$$

1.5

Using the convexity of max, then

$$\hat{R}(G) = \mathop{\mathbb{E}}_{\epsilon 1:n} \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g_i$$

Because max is a convex function, we can use Jensen's Inequality. Then,

$$\begin{aligned} \mathop{\mathbb{E}}_{\epsilon 1:n} \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g_i &\geq \max_{g \in G} \mathop{\mathbb{E}}_{\epsilon 1:n} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g_i \\ &= \max_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\epsilon 1:n} \epsilon_i g_i \end{aligned} \tag{2.4}$$

But $\mathbb{E}_{\epsilon 1:n} \epsilon_i = 0$. Therefore, $\hat{R}(G) \geq 0$, which is non-negative.

1.6

We start with the observation that $\hat{L}(\hat{f}_n) = \min_{f \in \mathcal{F}} \hat{L}(f)$. Then, using the property of concavity in Jensen's Inequality,

$$\begin{aligned} \mathbb{E}_{\mathcal{S}} \min_{f \in \mathcal{F}} \hat{L}(f) &\leq \min_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{S}} \hat{L}(f) \\ &= \min_{f \in \mathcal{F}} L(f) \\ &= L(f_{\mathcal{F}}) \end{aligned} \tag{2.5}$$

Problem 2 1.2

Let $\mathcal{X} = \mathcal{R}^d$, $\mathcal{Y} = [-M, M]$ and dataset of n points sampled from the distribution P, and $\mathcal{F} = f_1, ..., f_N$. Let l represent the square loss. Let us consider the following convergence problem:

$$P(\max_{f \in \mathcal{F}} |\mathbb{E}\, l(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} l(f(x_i, y_i)| \geq \epsilon)$$

In order to calculate the bound of the f that maximizes the difference, we need to consider every function $f \in \mathcal{F}$. This is the probability of the union of this inquality for each function in $\mathcal{F}$. Using the Union Bound Inequality and Hoeffding's Inequality, we obtain then obtain the upper bound,

$$\begin{aligned} P(\max_{f\in\mathcal{F}}|\mathbb{E}\, l(f(X),Y) - \frac{1}{n}\sum_{i=1}^{n} l(f(x_i,y_i)| \geq \epsilon) &= P(\bigcup_{k=1}^{N}|\mathbb{E}\, l(f_k(X),Y) - \frac{1}{n}\sum_{i=1}^{n} l(f_k(x_i),y_i)| \geq \epsilon) \\ &\leq \sum_{i=1}^{N} P(|\mathbb{E}\, l(f_k(X),Y) - \frac{1}{n}\sum_{i=1}^{n} l(f_k(x_i),y_i)| \geq \epsilon) \\ &= 2N\exp\{-\frac{2n\epsilon^2}{(C-M)^4}\} \end{aligned} \tag{2.6}$$

Where the bounds of the loss function are determined by the fact that it is square loss, so it has to be greater or equal to zero, and the upper bound is determined when $\sup_{x\in X}|f(x)| \leq C$ and $y = M$ such that $l = (C-M)^2$. 2.2

We begin by finding an expression for ϵ using the bound calculated above,

$$\begin{aligned} \delta &= 2N\exp\{-\frac{2n\epsilon^2}{(C-M)^4}\} \\ \log\frac{\delta}{2N} &= -\frac{2n\epsilon^2}{(C-M)^4} \\ \epsilon &= \sqrt{\frac{(C-M)^4}{2n}\log\frac{2N}{\delta}} \end{aligned} \tag{2.7}$$

If δ is the probability that the max of the differnce is greater than ϵ, then with probability $1-\delta$ we can say that the difference evaluated on the ERM solution will be less than or equal to ϵ.

$$L(\hat{f}_n) - \hat{L}(\hat{f}_n) \leq \epsilon(\delta,n,N)L(\hat{f}_n) \quad \leq \hat{L}(\hat{f}_n) + \epsilon(\delta,n,N) \tag{2.8}$$

Replacing ϵ for the one found above.

2.3

Let $\hat{f}_n$ be the ERM minimizer and $f_{\mathcal{F}}$ be the minimizer of the expected loss. Because of this, we note that $\hat{L}(f_{\mathcal{F}}) \geq \hat{L}(\hat{f}_n)$. Then,

$$\begin{aligned} L(\hat{f}_n) &= L(\hat{f}_n) - \hat{L}(f_{\mathcal{F}}) + \hat{L}(f_{\mathcal{F}}) \\ &\leq (\hat{L}(f_{\mathcal{F}}) - \hat{L}(\hat{f}_n) + L(\hat{f}_n) - \hat{L}(f_{\mathcal{F}}) + \hat{L}(f_{\mathcal{F}}) \\ &\leq (\hat{L}(f_{\mathcal{F}}) - L(f_{\mathcal{F}})) + (L(\hat{f}_n) - \hat{L}(\hat{f}_n)) + L(f_{\mathcal{F}}) \\ &\leq 2\max_{f\in\mathcal{F}}|L(f) - \hat{L}(f)| + L(f_{\mathcal{F}}) \end{aligned} \tag{2.9}$$

Therefore,

$$L(\hat{f}_n) - L(f_{\mathcal{F}}) \leq 2\max_{f\in\mathcal{F}}|L(f) - \hat{L}(f)|$$

With probability of $1-\delta$, then we can conclude that

$$L(\hat{f}_n) - L(f_{\mathcal{F}}) \leq 2\epsilon(\delta,n,N)$$