Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Bhaskar Mukhoty
Instructor: Purushottam Kar
Date: January 31, 2018

LECTURE
10

Radamacher Complexity

1 Introduction

In the last lecture we have seen that,

$$\mathbb{P}_S(\sup_{f\in\mathcal{F}} \left| er^l_{\mathcal{D}}[f] - er^l_S[f] \right| > 2 \cdot L \cdot R_n(\mathcal{F}) + \epsilon) \leq 2\exp(\frac{-n\epsilon^2}{2B^2}) \quad (1)$$

where $R_n(\mathcal{F})$ is the Radamacher complexity of the function class $\mathcal{F}$ evaluated over sample size n.

$$R_n(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{S,\hat{\epsilon}_i} \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i f(x_i) \right| \quad (2)$$

We are able to handle hinge loss, logistic loss, exponential loss, least square loss, ϵ-insensitive loss, using the Lipchitzness, but we shall not be able to handle, regularization or classification using l^{0-1}-loss which is not Lipchitz

We define empirical Radamacher average of a function class $\mathcal{F}$ w.r.t to sample set $S = \{x_1, x_2, ..., x_n\}$ and a set of radamacher random variables $\{\hat{\epsilon_1}, \hat{\epsilon_2}, ..., \hat{\epsilon_n}\}$ as

$$\mathop{\hat{R}}_{S,\hat{\epsilon}_i}(\mathcal{F}) \triangleq \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i f(x_i) \right| \quad (3)$$

Example 10.1. The emperical radamacher average for the set of linear classifiers, $\mathcal{W}$, would be:

$$\mathop{\hat{R}}_{S,\hat{\epsilon}_i}(\mathcal{W}) = \sup_{w\in\mathcal{W}} \left| \langle \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i, w\rangle \right|$$

Exercise 10.1. Show that $\mathbb{P}(\left| \mathop{\hat{R}}_{S,\hat{\epsilon}_i}(\mathcal{F}) - R_n(\mathcal{F})) \right| > \epsilon) \leq 2\exp(-\frac{n\epsilon^2}{2B^2})$ holds whenever $|f(x)| \leq B$
Hint: Prove that the function $g : (S, \hat{\epsilon}_i) \mapsto \mathop{\hat{R}}_{S,\hat{\epsilon}_i}(\mathcal{F})$ is stable.

There is an interesting relation between Gaussian and Radamacher complexity as below.

$$R_n(\mathcal{F}) \leq G_n(\mathcal{F}) \leq \ln(n) R_n(\mathcal{F})$$

where the Gaussian complexity $G_n(\mathcal{F})$, is defined by replacing Radamacher random variables by Gaussian random variables.

If we take expectation on equation (3) over $\{\epsilon_i\}$'s we get Radamacher complexity of the function class w.r.t. sample set S, i.e.

$$\underset{S}{R}(\mathcal{F}) \triangleq \underset{\hat{\epsilon}_i}{\mathbb{E}} \underset{S,\hat{\epsilon}_i}{\hat{R}}(\mathcal{F}) = \underset{\hat{\epsilon}_i}{\mathbb{E}} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right|$$

We can now observe that,

$$R_n(\mathcal{F}) = \underset{S}{\mathbb{E}} \underset{S}{R}(\mathcal{F}) \leq \sup_{S} \underset{S}{R}(\mathcal{F})$$

2 Massart's Finite Class Lemma (MFCL)

The lemma gives bound on $R_n(\mathcal{F})$ when the function class is finite, $|\mathcal{F}| < \infty$. Let, $S \in \mathcal{D}^n$, be a set of n-points from the feature space. For any $f \in \mathcal{F}$, we have, $f : (x_1, x_2, .., x_n) \mapsto (f(x_1), f(x_2), ..., f(x_n)) \in \mathbb{R}^n$. We define, a restriction with respect to S as,

$$\mathcal{A}_S = \{a \in \mathbb{R}^n : a = (f(x_1), f(x_2), ..., f(x_n)) \text{ for some } f \in \mathcal{F}\}$$

The MFCL lemma states that,

$$\underset{S}{R}(\mathcal{F}) \leq \frac{c}{n} \sqrt{2 \lg |\mathcal{A}_S|} \tag{4}$$

Since, each function in $f \in \mathcal{F}$ has only one evaluation for the set points $(x_1, x_2, .., x_n)$, we have, $|\mathcal{A}_S| \leq |\mathcal{F}|$

We can now re-define

$$\underset{S}{R}(\mathcal{F}) = \underset{\hat{\epsilon}_i}{\mathbb{E}} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| = \underset{\hat{\epsilon}_i}{\mathbb{E}} \sup_{a \in \mathcal{A}_S} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i a_i \right|$$

We would use Cramer-Chernoff to get a bound on the radacher average:

$$
\begin{aligned}
&\exp(s\mathop{\mathbb{E}}_{\hat{\epsilon}_i} \sup_{a\in\mathcal{A}_S} \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i a_i) \\
&\qquad \le \mathop{\mathbb{E}}_{\hat{\epsilon}_i} \exp(s \sup_{a\in\mathcal{A}_S} \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i a_i) && \text{[Jensen's Inequality]} \\
&\qquad = \mathop{\mathbb{E}}_{\hat{\epsilon}_i} \sup_{a\in\mathcal{A}_S} \exp(\frac{s}{n}\sum_{i=1}^{n} \hat{\epsilon}_i a_i) && \text{[Monotonicity of exponential]} \\
&\qquad = \mathop{\mathbb{E}}_{\hat{\epsilon}_i} \sup_{a\in\mathcal{A}_S} \prod_{i=1}^{n} \exp(\frac{s\hat{\epsilon}_i a_i}{n}) \\
&\qquad \le \mathop{\mathbb{E}}_{\hat{\epsilon}_i} \sum_{a\in\mathcal{A}_S} \prod_{i=1}^{n} \exp(\frac{s\hat{\epsilon}_i a_i}{n}) \\
&\qquad = \sum_{a\in\mathcal{A}_S} \prod_{i=1}^{n} \mathop{\mathbb{E}}_{\hat{\epsilon}_i} \exp(\frac{s\hat{\epsilon}_i a_i}{n}) \\
&\qquad = \sum_{a\in\mathcal{A}_S} \prod_{i=1}^{n} \exp(\frac{s^2 a_i^2}{2n^2}) && [\,-1 \le \hat{\epsilon}_i \le 1, \text{ Hoeffding's Lemma}] \\
&\qquad = \sum_{a\in\mathcal{A}_S} \exp(\frac{s^2}{2n^2}\sum_{i=1}^{n} a_i^2) \\
&\qquad = \sum_{a\in\mathcal{A}_S} \exp(\frac{s^2\left\|a\right\|_2^2}{2n^2}) \\
&\qquad = |\mathcal{A}_S| \exp(\frac{s^2c^2}{2n^2}) && \text{where, } \sup_{a\in\mathcal{A}_S} \left\|a\right\|_2 = c
\end{aligned}
$$

Taking log both sides we have,

$$
\mathop{R}_{S}(\mathcal{F}) \le \frac{1}{s}\lg|\mathcal{A}_S| + \frac{sc^2}{2n^2}
$$

Differentiating R.H.S w.r.t. s and setting it to zero, we have,

$$
\begin{aligned}
-\frac{1}{s^2}\lg|\mathcal{A}_S| + \frac{c^2}{2n^2} &= 0 \\
s^2 &= \frac{2n^2}{c^2}\lg|\mathcal{A}_S| \\
s &= \frac{n}{c}\sqrt{2\lg|\mathcal{A}_S|}
\end{aligned}
$$

So that,

$$
\begin{aligned}
\mathop{R}_{S}(\mathcal{F}) &\le \frac{1}{s}\lg|\mathcal{A}_S| + \frac{sc^2}{2n^2} \\
&= \frac{c}{n}\frac{\lg|\mathcal{A}_S|}{\sqrt{2\lg|\mathcal{A}_S|}} + \frac{c^2}{2n^2}\frac{n}{c}\sqrt{2\lg|\mathcal{A}_S|} \\
&= \frac{c}{n}\sqrt{\frac{\lg|\mathcal{A}_S|}{2}} + \frac{c}{n}\sqrt{\frac{\lg|\mathcal{A}_S|}{2}} \\
&= \frac{c}{n}\sqrt{2\lg|\mathcal{A}_S|}
\end{aligned}
$$

3 Applications of MFCL

3.1 Sparse Models

Let $\mathcal{F}$ be a linear class of sparse models, i.e.

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\|_0 \le s, \|w\|_\infty \le t\} \quad \text{where} \quad w, x \in \mathbb{R}^d, \quad \|x\|_\infty \le r$$

The above assumptions implies, $\|w\|_1 \le st$.
We would need Holder's inequality which states,
Now,

$$\begin{aligned}
\underset{S}{R}(\mathcal{F}) &= \mathbb{E}_{\hat{\epsilon}_i} \sup_{w\in\mathcal{W}} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i \langle w, x_i\rangle \right| \\
&= \mathbb{E}_{\hat{\epsilon}_i} \sup_{w\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i \langle w, x_i\rangle && [\text{since, } \hat{\epsilon}_i \in \{-1,1\} \iff -\hat{\epsilon}_i \in \{-1,1\}] \\
&= \mathbb{E}_{\hat{\epsilon}_i} \sup_{w\in\mathcal{W}} \langle w, \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i\rangle \\
&\le \mathbb{E}_{\hat{\epsilon}_i} \sup_{w\in\mathcal{W}} \left\| w \circ (\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i) \right\|_1 && [\circ \text{ denotes element-wise or Hadamard product}] \\
&\le \mathbb{E}_{\hat{\epsilon}_i} \sup_{w\in\mathcal{W}} \|w\|_1 \left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i \right\|_\infty && [\text{Hölder's Inequality}] \\
&\le st\, \mathbb{E}_{\hat{\epsilon}_i} \left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i \right\|_\infty \\
&= st\, \mathbb{E}_{\hat{\epsilon}_i} \sup_{j\in[d]} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i^j \right|
\end{aligned}$$

Consider the set $A = \{j \in [d] : (x_1^j, x_2^j, ..., x_n^j)\}$, so that,

$$\begin{aligned}
R(A) &= \mathbb{E}_{\hat{\epsilon}_i} \sup_{j\in[d]} \left| \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i x_i^j \right| \\
&\le \frac{c}{n}\sqrt{2\lg|\mathcal{A}|} \\
&= \frac{r\sqrt{n}}{n}\sqrt{2\lg d}
\end{aligned}$$

where $c = \left\|x^j\right\|_2 \le \sqrt{nr^2} = r\sqrt{n}$ and $|\mathcal{A}| = d$
Hence,

$$\underset{S}{R}(\mathcal{F}) \le srt\sqrt{\frac{2\lg d}{n}} \tag{5}$$

3.2 Covering Number and Radamacher Average

Let $\mathcal{F}, \mathcal{C}_\epsilon$ be function classes defined over the set $\mathcal{X}$, where $\mathcal{C}_\epsilon$ is a ϵ-covered of $\mathcal{F}$. That is, $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{C}_\epsilon$, such that $\forall x \in \mathcal{X}$, $|f(x) - g(x)| \le \epsilon$.

First we will show,

$$R_n(\mathcal{F}) \leq \epsilon + R_n(\mathcal{C}_\epsilon) \tag{6}$$

Let, $\epsilon_i \in \{-1, 1\}$ is radamacher random variable, then

$$\begin{aligned}
\left|\sum_{i=1}^{n} \epsilon_i f(x_i)\right| &= \left|\sum_{i=1}^{n} \epsilon_i f(x_i) - \sum_{i=1}^{n} \epsilon_i g(x_i) + \sum_{i=1}^{n} \epsilon_i g(x_i)\right| \\
&\leq \left|\sum_{i=1}^{n} \epsilon_i f(x_i) - \sum_{i=1}^{n} \epsilon_i g(x_i)\right| + \left|\sum_{i=1}^{n} \epsilon_i g(x_i)\right| && \text{[Triangle inequality]} \\
&= \left|\sum_{i=1}^{n} \epsilon_i (f(x_i) - g(x_i))\right| + \left|\sum_{i=1}^{n} \epsilon_i g(x_i)\right| \\
&\leq \sum_{i=1}^{n} |\epsilon_i (f(x_i) - g(x_i))| + \left|\sum_{i=1}^{n} \epsilon_i g(x_i)\right| && \text{[Triangle inequality]} \\
&= \sum_{i=1}^{n} |\epsilon_i| \, |(f(x_i) - g(x_i))| + \left|\sum_{i=1}^{n} \epsilon_i g(x_i)\right| \\
&= n\epsilon + \left|\sum_{i=1}^{n} \epsilon_i g(x_i)\right| && [|\epsilon_i| = 1, |f(x) - g(x)| \leq \epsilon]
\end{aligned}$$

So that,

$$\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)\right| \leq \epsilon + \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(x_i)\right| \\
\text{or,} \quad & \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)\right| \leq \epsilon + \sup_{g \in \mathcal{C}_\epsilon} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(x_i)\right| \\
\text{or,} \quad & \mathop{\mathbb{E}}_{S,\epsilon_i} \sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)\right| \leq \epsilon + \mathop{\mathbb{E}}_{S,\epsilon_i} \sup_{g \in \mathcal{C}_\epsilon} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(x_i)\right| \quad [\text{ if } X \leq Y \text{ then } \mathbb{E}[X] \leq \mathbb{E}[Y]]
\end{aligned}$$

which implies equation (6).

Let, $|f(x)| \leq B$, which implies $|g(x)| \leq B$, as we can always find, $\mathcal{C}_\epsilon \subseteq \mathcal{F}$. Since, $\underset{S}{R}(\mathcal{C}_\epsilon) \leq \frac{B}{n}\sqrt{2 \lg |\mathcal{C}_\epsilon|}$, we have, $R_n(\mathcal{C}_\epsilon) = \underset{S}{\mathbb{E}}\underset{S}{R}(\mathcal{C}_\epsilon) \leq \frac{B}{n}\sqrt{2 \lg |\mathcal{C}_\epsilon|}$.

So that,

$$\begin{aligned}
R_n(\mathcal{F}) &\leq \epsilon + R_n(\mathcal{C}_\epsilon) \\
&\leq \epsilon + \frac{B}{n}\sqrt{2 \lg |\mathcal{C}_\epsilon|} \\
&\leq \epsilon + \frac{B}{n}\sqrt{2d \lg(1 + \frac{2B}{\epsilon})} && \text{[for linear models of 2-norm at most } B\text{]}
\end{aligned}$$

In general we can write,

$$R_n(\mathcal{F}) \leq \inf_{\alpha}\{\alpha + \frac{B}{n}\sqrt{2\lg N_{\mathcal{X}}(\mathcal{F},\alpha)}\} \tag{7}$$

3.3 Classification

Let, $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$, is evaluated using l^{0-1} loss, which is non-Liptchitz. Then we can prove the following using Mc'Diarmid inequality and boundedness of loss-function.

$$\mathcal{P}_S(\sup_{f\in\mathcal{F}} \left|er_D^{0-1}[f] - er_S^{0-1}[f]\right| > 2R_n(l^{0-1}\circ\mathcal{F}) + \epsilon) \leq 2\exp(\frac{-n\epsilon^2}{2}) \tag{8}$$

Although, l^{0-1} is non-lipchitz, we can have the following inequality:

$$R_n(l^{0-1}\circ\mathcal{F}) \leq \frac{1}{2}R_n(\mathcal{F})$$

Definition 10.1. (Restriction of $\mathcal{F}$ to $\mathcal{A}$): Let $\mathcal{F} \subseteq \{-1,1\}^{\mathcal{X}}$ be a function class and $\mathcal{A} = \{x_1, x_2, ...x_n\} \subset \mathcal{X}$. The restriction of $\mathcal{F}$ to $\mathcal{A}$ is the set of possible functions in $\mathcal{F}$, from $\mathcal{A}$ to $\{-1,1\}$. That is,

$$\mathcal{F}_{\mathcal{A}} = \{(f(x_1), f(x_2), .., f(x_n)) : f \in \mathcal{F}\}$$

Each function in the restriction is a vector $\{-1,1\}^n$Shalev-Shwartz and Ben-David (2014)

Definition 10.2. (Shattering:) A function class $\mathcal{F}$ shatters a finite set $\mathcal{A} \subset \mathcal{X}$, if the restriction $\mathcal{F}_{\mathcal{A}}$, has all the possible function from $\mathcal{A}$ to $\{-1,1\}$. That is, $|\mathcal{F}_{\mathcal{A}}| = 2^{|\mathcal{A}|}$

Definition 10.3. (Growth Function:) The growth of a function class $\mathcal{F}$, denoted by $\Pi_n(\mathcal{F}) : \mathbb{N} \to \mathbb{N}$, is defined as:

$$\Pi_n(\mathcal{F}) = \max_{\mathcal{A}\subset\mathcal{X}, |\mathcal{A}|=n} |\mathcal{F}_{\mathcal{A}}| \tag{9}$$

We would always have, $\Pi_n(\mathcal{F}) \leq 2^n$
By definition we can see,

$$R_n(\mathcal{F}) \leq \tfrac{c}{n}\sqrt{2\lg\Pi_n(\mathcal{F})} = \sqrt{\tfrac{2\lg\Pi_n(\mathcal{F})}{n}} \qquad \text{since, } c = \max_{a\in\mathcal{A}} \|a\|_2 = \sqrt{n}$$

In the next lecture, we would get a useful bound on the radamacher complexity of $\mathcal{F}$ using upper bound on the growth function.

References

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.