Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Archit Sharma
Instructor: Purushottam Kar
Date: January 22, 2018

LECTURE 7

Covering Number Bound

1 Introduction

In Garg and Kar (2018), we showed uniform convergence for a finite hypothesis class on a toy binary classification. However, it is non-trivial to extend the notion of uniform convergence (if it exists) to an infinite hypothesis class, as is the case in real regression we are dealing with.

First, we introduce the *Hoeffding's Inequality*, which is a generalization of *Chernoff Bound* to bounded random variables. This immediately gives us pointwise convergence for the real regression, where we assume both our parameter and output space to be bounded. Next, we introduce the notion of *Covering Numbers* and give an intuition as to how we will obtain the uniform convergence for the real regression hypothesis space using it (albeit in presence of some additional assumptions). In this lecture, we will formally show uniform convergence using *covering numbers* and also introduce the *McDiarmid's Inequality.*

2 Covering Number Bound

Recall our setup for real regression. The input space is represented by $\mathcal{X} \subseteq \mathbb{R}^d$, such that $\forall x \in \mathcal{X}, \|x\|_2 \leq 1$. The output space is denoted by $\mathcal{Y} \in [-B, B]$. The hypothesis space $\mathcal{F}$ is defined as $\{x \mapsto \langle w, x\rangle : x \in \mathcal{X}, \|w\|_2 \leq B\}$. We use the squared loss function $l_{sq}(\hat{y}, y) = (\hat{y} - y)^2$ which is $4B$-Lipschitz (Verify!). Next, we formalize the notions of "covering" and "packing" the sets.

Definition 7.1 (ϵ-cover)**.** Given a set $\mathcal{X}$ with a norm $\|.\|$ defined over it, a set $\mathcal{C} \subseteq \mathcal{X}$ is said to be an ϵ-cover for $\mathcal{X}$ if $\forall x \in \mathcal{X}, \exists c_x \in \mathcal{C}$ such that $\|x - c_x\| \leq \epsilon$.

Corollary 7.1. $\mathcal{X} \subseteq \bigcup\limits_{c \in \mathcal{C}} \mathcal{B}(c, \epsilon)$ where $\mathcal{B}(c, \epsilon) \triangleq \{x \in \mathcal{X} : \|x - c\| < \epsilon\}$

Mathematical jargon aside, a set $\mathcal{C}$ is said to be an ϵ-cover when the set of ϵ-radius balls centred on all points in $\mathcal{C}$ "cover" the set $\mathcal{X}$, that is, every point $x \in \mathcal{X}$ needs to belong to *atleast* one such ball for the set $\mathcal{X}$ to be covered.

Definition 7.2 (ϵ-packing)**.** A set $\mathcal{C} \subseteq \mathcal{X}$ is said to be an ϵ-packing if $\forall c, c' \in \mathcal{C}$ such that $c \neq c'$, $\|c - c'\| > \epsilon$.

Corollary 7.2. $\mathcal{B}(c, \frac{\epsilon}{2}) \bigcap \mathcal{B}(c', \frac{\epsilon}{2}) = \emptyset \ \forall c, c' \in \mathcal{C}$ such that $c \neq c'$.

An ϵ-packing can be understood to be a set such that any two points in the set are at least ϵ distance away (under the previously defined norm over the set). For example, a single point

trivially constitutes a ϵ-packing. Now, the $\frac{\epsilon}{2}$ balls centred around the points in the ϵ-packing are *not allowed to overlap.*

Let's relate it to the discussion on real regression. We introduced the notion of covering numbers to get a bound on the "representative" hypothesis from $\mathcal{F}$. These representatives are such that $\forall w' \in \mathcal{F}$, there exists a w among the "representatives" such that $\|w - w'\|_2 \leq \epsilon$. Since, the squared loss function is Lipschitz, the loss incurred on the hypothesis will not be "very different" for the same set of samples. This crucially establishes the following fact: If the difference between the sample error and L-risk is "small" for representative hypotheses, it will be "small" for the entire hypothesis space. Thus, proving uniform convergence on these finitely many representative hypothesis (which is not hard once the bound on these representatives is known), will prove uniform convergence on the entire hypothesis class. We will show the complete proof formally, but this discussion should highlight the trajectory for the rest of the proof.

This defines two tasks: 1) Get a finite upper bound on the number of "representative hypotheses" and 2) Show uniform convergence for our hypothesis space using these "representative". From the above explanation of "representative" hypothesis, it is clear that the we are looking for an ϵ-cover for $\mathcal{F}$. Further, for tight bounds on uniform convergence, we will want to minimize the number of "representatives", that is reduce the cardinality of ϵ-cover set. We obtain a bound on the minimal ϵ-cover indirectly, using the concept of ϵ-packing. Looking at their definitions (in particular the corollaries), the notions of ϵ-cover and ϵ-packing look contrary. The following definitions and theorem will formalize this intuition.

Definition 7.3 (Minimal ϵ-cover)**.** A set $\mathcal{C} \subseteq \mathcal{X}$ is said to be the minimal ϵ-cover of $\mathcal{X}$ if it is an ϵ-cover and $|\mathcal{C}| \leq |\mathcal{C}'|$ for all $\mathcal{C}'$ which are ϵ-covers of $\mathcal{X}$.

Definition 7.4 (Maximal ϵ-packing)**.** A set $\mathcal{C} \subseteq \mathcal{X}$ is said to be a maximal ϵ-packing if $\mathcal{C}$ is an ϵ-packing and $|\mathcal{C}| \geq |\mathcal{C}'|$ for all $\mathcal{C}'$ which are ϵ-packings with respect to $\mathcal{X}$

Theorem 7.3. A maximal $\frac{\epsilon}{2}$-packing is always an ϵ-cover.

Proof. Suppose a set $\mathcal{C}$ is an $\frac{\epsilon}{2}$-packing but not an ϵ-cover.
$\Rightarrow \exists$ a point $x_0 \in \mathcal{X}$ such that $\min_{c \in \mathcal{C}} \|x_0 - c\| > \epsilon$
$\Rightarrow \mathcal{C} \bigcup \{x_0\}$ is an $\frac{\epsilon}{2}$ packing $\Rightarrow$ Contradiction! □

Thus, estimating the number of points in the maximal $\frac{\epsilon}{2}$-packing will provide the upper bound on the cardinality of the minimal ϵ-cover. Let $\mathcal{N}_{\mathcal{X}}(\epsilon)$ denote the cardinality of minimal ϵ-cover and $\mathcal{P}_{\mathcal{X}}(\epsilon)$ denote the cardinality of maximal ϵ-packing on the set $\mathcal{X}$. From the theorem above, we get,

$$\mathcal{N}_{\mathcal{X}}(\epsilon) \leq \mathcal{P}_{\mathcal{X}}(\tfrac{\epsilon}{2})$$

From the definition of ϵ-packing, the balls centred around the points in the set are not allowed to overlap. Therefore,

$$\mathcal{P}_{\mathcal{X}}(\epsilon) Vol(\tfrac{\epsilon}{2}) \leq Vol(B + \tfrac{\epsilon}{2})$$

Here we assume our $\mathcal{X}$ to be a subset of a ball of radius B. We have allowed the packed volume to go upto a ball of radius $B + \frac{\epsilon}{2}$ because points at the boundary can have their balls outside the volume enclosed by the hypothesis space. The above inequality leads to

$$\mathcal{P}_{\mathcal{X}}(\epsilon) \leq \frac{Vol(B+\frac{\epsilon}{2})}{Vol(\frac{\epsilon}{2})} = \left(1 + \frac{2B}{\epsilon}\right)^d$$

where d is the dimensionality of the hypothesis space. Now,

$$\mathcal{N}_{\mathcal{X}}(\epsilon) \leq \mathcal{P}_{\mathcal{X}}(\tfrac{\epsilon}{2}) \leq \left(1 + \tfrac{4B}{\epsilon}\right)^d$$

This is the *Covering Number Bound* for our hypothesis space. Armed with this, we move to the next objective and show a formal proof of Uniform Convergence in the next section.

Remark 7.1. The covering number bound is a measure of complexity of the hypothesis space $\mathcal{F}$, much like other measures of complexity like Rademacher Complexity, VC dimension (to be introduced later). The covering number bound is polynomial in B, exponential in the dimensionality of data d. Therefore, more expressive a function class, larger the covering number is. Computation of covering number is NP-Hard, but, we are only interested in computing the bound.

3 Uniform Convergence for Real Regression

For a rigorous treatment of this topic, refer to Cucker and Smale (2002). We have already shown that if $|f(x) - f'(x)| \leq \epsilon$ for all $x \in \mathcal{X}$, then $\forall S \in (\mathcal{X} \times \mathcal{Y})^n$

$$\left| |er^l_S[f] - er^l_{\mathcal{D}}[f]| - |er^l_S[f] - er^l_{\mathcal{D}}[f']| \right| \leq 8B\epsilon$$

The above statement also implies that if $|er^l_S[f] - er^l_{\mathcal{D}}[f]| \leq \epsilon \Rightarrow |er^l_S[f] - er^l_{\mathcal{D}}[f']| \leq (8B+1)\epsilon$. For uniform convergence, we have to show that

$$\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right| \geq \epsilon\right] \leq \delta$$

Let $\mathcal{C} = \{f_1, \dots f_h\}$ represent the minimal $\frac{\epsilon}{8B+1}$-cover of $\mathcal{F}$, where $h = \mathcal{N}_{\mathcal{F}}(\frac{\epsilon}{8B+1})$ and let $\mathcal{F}_i$ represents the $\frac{\epsilon}{8B+1}$ radius ball centred on $f_i \in \mathcal{F}$.

Lemma 7.4. $\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right| \geq \epsilon\right] \leq \sum_{i=1}^{h} \mathbb{P}\left[\sup_{f \in \mathcal{F}_i} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right| \geq \epsilon\right]$

Proof. Uses the Union Bound and the fact that $\bigcup_{i=1}^{h} \mathcal{F}_i$ covers $\mathcal{F}$. □

Lemma 7.5. $\mathbb{P}\left[\sup_{f \in \mathcal{F}_i} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right| \geq \epsilon\right] \leq \mathbb{P}\left[\left|er^l_{\mathcal{D}}[f_i] - er^l_S[f_i]\right| \geq \frac{\epsilon}{8B+1}\right]$

Proof. Follows from the discussion above. Let $\hat{f} = \sup_{f \in \mathcal{F}_i} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right|$. Since, $f, \hat{f} \in \mathcal{F}_i \Rightarrow |f(x) - \hat{f}(x)| \leq \epsilon$. Therefore, if $\left|er^l_{\mathcal{D}}[f_i] - er^l_S[f_i]\right| \leq \frac{\epsilon}{8B+1} \Rightarrow \left|er^l_{\mathcal{D}}[\hat{f}] - er^l_S[\hat{f}]\right| \leq \epsilon \Rightarrow \mathbb{P}\left[\left|er^l_{\mathcal{D}}[\hat{f}] - er^l_S[\hat{f}]\right| \geq \epsilon\right] \leq \mathbb{P}\left[\left|er^l_{\mathcal{D}}[f_i] - er^l_S[f_i]\right| \geq \frac{\epsilon}{8B+1}\right]$. □

Theorem 7.6. For a $4B$-Lipschitz loss function $l \in [-B, B]$, $\mathbb{P}\left[\sup_{f \in \mathcal{F}} \left|er^l_{\mathcal{D}}[f] - er^l_S[f]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{32B^4(8B+1)^2} + d\log\left(1 + \frac{4B(8B+1)}{\epsilon}\right)\right)$

Proof. Follows directly from the lemmas and the pointwise convergence of the hypothesis class. Use the covering bound to simplify the expressiong, that is $\mathcal{N}_{\mathcal{F}}(\frac{\epsilon}{8B+1}) \leq \left(1 + \frac{4B(8B+1)}{\epsilon}\right)^d$. Note the neighbourhood size is $\frac{\epsilon}{8B+1}$, and not ϵ, because we are considering representative hypothesis with error $\frac{\epsilon}{8B+1}$. □

Remark 7.2. This bound applies to a lot of machine learning problems, that is whenever the loss is Lipschitz and bounded. This is the case for Hinge loss, Logistic regression, ϵ-insensitive loss as long as the hypothesis space is bounded. Note that this does not apply to $0-1$ error as the loss function is not Lipschitz.

Remark 7.3. We have shown uniform convergence for an infinite hypothesis class by using *covering numbers*. However, the analysis is clearly messy, involving tedious constants. Also, we have now introduced a quartic dependence on B and a dependence on dimensionality d, both of which are not desirable. An analysis based on Rademacher Complexity for the same bound will be less messier, and also make the bound dimensionality independent.

4 McDiarmid's Inequality

To conduct Rademacher Complexity based analysis of the bound, we need to add more tools to our inventory. The first such tool is "McDiarmid's Inequality", which is a powerful concentration inequality.

Definition 7.5 (c-stability)**.** A function $f : \mathcal{X}^m \mapsto \mathbb{R}$ is said to be c-stable if $\forall x_1, x_2, \dots x_m \in \mathcal{X}, j \in [m], x_i' \in \mathcal{X}$ if

$$|f(x_1, x_2, \dots x_i, \dots x_m) - f(x_1, x_2, \dots x_i', \dots x_m)| \leq c$$

As an example of a c-stable function, consider $f(x_1, \dots x_m) = \frac{1}{n}\sum_{i=1}^m x_i$ where $x_i \in [a, b]$.

$$|f(x_1, x_2, \dots x_i, \dots x_m) - f(x_1, x_2, \dots x_i', \dots x_m)| = \frac{1}{m}|x_i - x_i'| \leq \frac{b-a}{m} = c$$

Theorem 7.7 (McDiarmid's Inequality)**.** Let $f : \mathcal{X}^m \mapsto \mathbb{R}$ be a c-stable function, then for any $\mathcal{D}$ over $\mathcal{X}$,

$$\mathbb{P}\left[\left|f(x_1, \dots, x_m) - \mathbb{E}\left[f\right]\right| > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{mc^2}\right)$$

The proof for this inequality will be provided in the next lecture.

Remark 7.4. It is clear that the McDiarmid's is a much more general inequality when compared to Hoeffding's Inequality, or other inequalities like Chernoff and Chebyshev's inequalities. In fact, Hoeffding's can be derived as a corollary by considering the empirical average as the function (considered in the example for c-stability). One of the important aspect of this inequality is that it does not constrain to empirical averages, like Hoeffding's and Chernoff do. c-stable functions are easier to manipulate than using empirical averages. Consider the random variable $X_S = \sup_{f \in \mathcal{F}} \left\{er_{\mathcal{D}}^l[f] - er_S^l[f]\right\}$. This random variable maps a $S \in \mathcal{X}^N \mapsto \mathbb{R}$, and is also c-stable. Note, that is exactly the random variable which needed to bound for showing uniform convergence. McDiarmid's Inequality provides a route for an analysis where X_S like random variables are upper bounded.

References

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Abhibhav Garg and Purushottam Kar. Topics in Learning Theory, Lecture 4. Technical report, Indian Institute of Technology, Kanpur, 2018.