36-705: Intermediate Statistics **Fall 2019**

Lecture 10: September 18

Lecturer: Siva Balakrishnan

In the last lecture we discussed the relationship between the uniform convergence of empirical probabilities to true ones over collections of sets and the VC dimension of the collection of sets. Today we return to the uniform convergence over classes of functions, and relate this to the *Rademacher complexity,* of the collection of functions.

10.1 Empirical process

As a reminder, the setup for today is that we have a collection of functions $\mathcal{F}$, we observe samples $X_1, \ldots, X_n \sim P$ for some distribution P and we are interested in (upper bounding) the quantity:

$$\Delta(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

10.2 Rademacher complexity

Unlike the VC dimension, in the definition of the Rademacher complexity we do not maximize over the locations of points, i.e. in some sense it is not a worst case measure of complexity. In order to define the Rademacher complexity, we first suppose that we have a fixed collection $\{x_1, \ldots, x_n\}$ of points.

We let $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$ denote a collection of n Rademacher random variables, i.e. they take the values $\{+1, -1\}$ with equal probabilities. In this case, we can define the *empirical* Rademacher complexity as:

$$\mathcal{R}(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right].$$

When we think of $\{x_1, \ldots, x_n\}$ as a random sample then the empirical Rademacher complexity is a random variable. We define the Rademacher complexity of the class $\mathcal{F}$ as the expectation of this quantity, i.e.

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right].$$

Just intuitively, we should think about when the Rademacher complexity is large, and when it decays to 0. The Rademacher complexity is measuring the maximum absolute covariance between $\{f(X_1), \ldots, f(X_n)\}$ and a vector of random signs $\{\epsilon_1, \ldots, \epsilon_n\}$.

Intuitively, we think of a class $\mathcal{F}$ as too large if for many random sign vectors we can find a function in $\mathcal{F}$ that is strongly correlated with the random sign vectors.

The main utility of the Rademacher complexity is that it upper bounds the quantity $\Delta(\mathcal{F})$ that we care about.

Rademacher Theorem:

$$\mathbb{E}[\Delta(\mathcal{F})] \leq 2\mathcal{R}(\mathcal{F}).$$

This theorem again might not appear to be so useful since we still need to understand the Rademacher complexity. It turns out that the Rademacher complexity is relatively easy to upper bound in terms of more geometric measures of the function class $\mathcal{F}$ (these are things like covering numbers or bracketing numbers of $\mathcal{F}$). This is analogous to how VC theory gave us a way to go from the uniform convergence question to a combinatorial property of the collection of sets. You will see these in more detail in 702.

Proof: At a high-level the proof will resemble what we did in proving Hoeffding's inequality. We will introduce a ghost sample, and symmetrize the empirical process. Concretely, let $\{Y_1, \ldots, Y_n\}$ be an independent identically distributed sample. Then,

$$\begin{aligned}
\mathbb{E}[\Delta(\mathcal{F})] &= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f] \right| \right] \\
&= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_i} f(Y_i) \right| \right] \\
&= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_Y \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right| \right] \\
&\leq \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \mathbb{E}_Y \left| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right| \right] \\
&\leq \mathbb{E}_{X,Y} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right| \right]
\end{aligned}$$

We note that the distribution of the difference $f(X_i) - f(Y_i)$ is the same as the distribution

of $\epsilon_i(f(X_i) - f(Y_i))$ so we obtain,

$$\begin{aligned}\mathbb{E}[\Delta(\mathcal{F})] &\leq \mathbb{E}_{X,Y,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[f(X_i) - f(Y_i)]\right|\right] \\ &\leq 2\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right] \\ &= 2\mathcal{R}(\mathcal{F}),\end{aligned}$$

which gives us the Rademacher theorem.

If the function class is bounded, i.e. for every $f \in \mathcal{F}$ we have that $\|f\|_\infty \leq b$, then the empirical process $\Delta(\mathcal{F})$ is sharply concentrated around its mean, i.e.

$$\mathbb{P}(|\Delta(\mathcal{F}) - \mathbb{E}[\Delta(\mathcal{F})]| \geq t) \leq 2\exp(-nt^2/(2b^2)).$$

This inequality is a consequence of McDiarmid's inequality we studied previously. We won't go through this argument but it is a great exercise (HW4?).

Putting this inequality together with the upper bound on the mean we obtain that for a bounded class $\mathcal{F}$ with probability at least $1 - \delta$,

$$\Delta(\mathcal{F}) \leq 2\mathcal{R}(\mathcal{F}) + b\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Noting that we obtain the concentration bound rather easily, the quantity that is often difficult to deal with is $\mathcal{R}(\mathcal{F})$. We'll consider a few examples and leave the rest to 702.

10.3 Rademacher Complexity of a Finite Class

Suppose that we have a finite collection of functions $\mathcal{F} = \{f_1, \ldots, f_N\}$, which are bounded i.e. $\|f_i\|_\infty \leq b$ then we have the following bound on the Rademacher complexity.

Finite Class Bound: The Rademacher complexity for a finite class,

$$\mathcal{R}(\mathcal{F}) \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

Note that in this case it would in fact be more direct to work with the empirical process, and upper bound that directly. This is not usually the case.

Proof: Define,

$$\Theta := \mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right].$$

For convenience let us augment the class $\mathcal{F}$ with the negative of every function, i.e. we take $\widetilde{\mathcal{F}} = \mathcal{F} \cup (-\mathcal{F})$, so that there are now $2N$ functions. Then,

$$\Theta \leq \mathbb{E}_{X,\epsilon}\left[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right].$$

Note that,

$$\begin{aligned}
\exp(t\Theta) &\leq \exp\left(t\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right]\right) \\
&\leq \mathbb{E}_{X,\epsilon}\exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right]\right) \\
&\leq \mathbb{E}_{X,\epsilon}\sum_{j=1}^{2N}\prod_{i=1}^{n}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right) \\
&= \sum_{j=1}^{2N}\prod_{i=1}^{n}\mathbb{E}_{X,\epsilon}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right).
\end{aligned}$$

Since $\|f_j\|_\infty \leq b$ we can use the argument we used in the proof of Hoeffding's inequality to obtain that,

$$\exp(t\Theta) \leq 2N\exp\left(\frac{4t^2b^2}{n}\right),$$

so we obtain that,

$$\Theta \leq \frac{\log(2N)}{t} + \frac{4tb^2}{n},$$

where t is a free parameter that is > 0. Choosing, $t = \sqrt{n\log(2N)/(4b^2)}$ we obtain,

$$\Theta \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

10.4 Using the Rademacher Theorem to obtain the VC theorem

The Rademacher theorem in a very straightforward way implies the VC theorem. We'll sketch the proof here. Our class of functions just corresponds to the indicators arising from the set system. These functions are upper bounded by $b = 1$. We can get a high-probability statement as in the initial section so we only need to deal with $\mathcal{R}(\mathcal{F})$.

We follow an identical argument to the one we did in the previous section,

$$\exp(t\Theta) \leq \mathbb{E}_{X,\epsilon} \exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right]\right),$$

where the class $\widetilde{\mathcal{F}}$ just contains the set indicators and their negations.

The key point is to note here is the following: suppose we think of the vectors $(f(X_1), \ldots, f(X_n))$ for each function in $\widetilde{\mathcal{F}}$ and ask how many different such vectors are there? Each set in $\mathcal{A}$ picks out some subset of the points (and assigns them the value $+1$). Even though there are possibly infinitely many sets in $\mathcal{A}$ there are at most only twice (because we included the negations) the shattering number of different vectors.

The shattering number is precisely the (maximum) number of different vectors $(f(X_1), \ldots, f(X_n))$ we can induce using our collection of sets.

With this insight in hand we can just repeat the previous argument to conclude that,

$$\Theta \leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}},$$

and putting this together with the high-probability bound from before we have that with probability at least $1-\delta$,

$$\begin{aligned}\Theta &\leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}} \\ &\leq \sqrt{\frac{4\log(4s(\mathcal{A}, n)/\delta)}{n}},\end{aligned}$$

which is precisely the VC theorem (again always ignore constants).