Indian Institute of Technology Kanpur
CS777 Topics in Learning Theory

Scribe: Nishit Asnani
Instructor: Purushottam Kar
Date: January 29th, 2018

LECTURE
9

Symmetrization

1 Introduction

"Symmetry is what we see at a glance; based on the fact that there is no reason for any difference..."

- Blaise Pascal, Pensées

In our quest to establish uniform convergence for linear regression for an infinite function class, we had looked at the method of covering numbers to get a concentration inequality for the difference between the training and test error. The dependence of the error probability bound on the exponential of the data dimensionality though, provided us with a formidable challenge: how do we prove a non-trivial concentration bound for the case of high-dimensional data?

In the previous lecture, we looked at McDiarmid's inequality, which provides a neat guarantee of concentration for 'stable' functions. This would be the first tool in our armory to break down the 'curse' that high dimensional inputs put on us. In this lecture, we will develop a few more tools to guide us on our path to salvation, i.e. a form of uniform convergence that is independent of data dimensionality. On the way, we would introduce the notion of Rademacher complexity of a function class, which will come in handy for our analysis.

Symmetry is nature's way of portraying beauty to us. It is also one of the most profound tools used by mathematicians and physicists to analyze certain phenomena. The best part about it though, is that it is so intuitive that one can explain an argument of symmetry even to a novice. Symmetry would be crucially exploited in this lecture to arrive at the Rademacher complexity bound for the expected error difference.

2 Motivation

Our setup for real regression is as follows. The input space is represented by $\mathcal{X} \subseteq \mathbb{R}^m$, and the labels are given by $\mathcal{Y} \in [-b, b]$. $S \in (\mathcal{X} \times \mathcal{Y})^n$ is the training dataset.

For a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and a bounded loss function $l \in [0, B]$ (where $B = 4b^2$ for squared loss function l), we showed that $g : S \mapsto \sup_{f \in \mathcal{F}} \{er_S^l[f] - er_{\mathcal{D}}^l[f]\}$ is $\mathcal{O}\left(\frac{B}{n}\right)$-stable. Applying McDiarmid's inequality to g_S then gives us,

$$\mathop{\mathbb{P}}_{S}\Big[g_S > \mathbb{E}\left[g_S\right] + \epsilon\Big] \leq exp\Big(\frac{-n\epsilon^2}{2B^2}\Big) \tag{1}$$

Remark 9.1. Notice that the above result doesn't depend on the function class $\mathcal{F}$.

Although 1 gives us a concentration guarantee on the error function g_S around its expectation, it does not provide any bounds on $\mathbb{E}[g_S]$. This implies that even though we know that g_S would lie close to its expected value, if the latter is large, we would run into trouble as g_S would also be large with high probability. This would ensure that uniform convergence has a hard time (empirical risk and l-risk could be quite distant for at least some of the functions in $\mathcal{F}$).

In order to alleviate this problem, we try to bound $\mathbb{E}[g_S]$. For the given setting, $|g_S| \leq B$ trivially, and therefore, $\mathbb{E}[g_S] \leq B$. Ideally, for function classes that satisfy uniform convergence, we would want $\mathbb{E}[g_S] \leq B/\sqrt{n}$, B/n or something similar, that scales down with increase in the size of S.

We would now look at some of the tools that we would need to bound $\mathbb{E}[g_S]$.

3 Tools Required

This is what the rest of our armory consists of:

- $\forall f \in \mathcal{F}, \mathop{\mathbb{E}}_{\tilde{S}}\left[er^l_{\tilde{S}}[f]\right] = er^l_{\mathcal{D}}[f]$ for a set $\tilde{S}$ sampled from $\mathcal{D}^n$. This has a clean proof using linearity of expectation, and is left as an exercise to the reader.
- Jensen's inequality: For a convex function $h : \mathcal{Z} \mapsto \mathbb{R}$, we have $\mathop{\mathbb{E}}_{z}[h(z)] \geq h(\mathbb{E}[z])$. A corresponding inequality holds for concave functions as well, with the sign reversed.
- Setting $e_f(S) = er^l_S[f] - er^l_{\mathcal{D}}[f]$, $\sup_{f\in\mathcal{F}}\{e_f(S)\}$ is a convex function.
- For any two functions $e^1_f(S)$ and $e^2_f(S)$ in S, we have,
  $\sup_{f\in\mathcal{F}}\{e^1_f(S) + e^2_f(S)\} \leq \sup_{f\in\mathcal{F}}\{e^1_f(S)\} + \sup_{f\in\mathcal{F}}\{e^2_f(S)\}$.

  For sequences of real numbers $\{a_i\}$ and $\{b_i\}$, $i \in [n]$, $\max_{i\leq n}\{a_i + b_i\} \leq \max_{i\leq n}\{a_i\} + \max_{i\leq n}\{b_i\}$. We can substitute a_i by $e^1_f(S)$ and b_i by $e^2_f(S)$, and run i over all functions f in $\mathcal{F}$ to get the desired result.
- The concept of symmetrization, which is explained in the following subsection.

Exercise 9.1. Prove the third point by observing that supremum of a family of convex functions is convex. Define $e_f(S_1 + S_2)$ suitably to arrive at the proof.

3.1 Symmetrization

Let's suppose we have a bag of n balls, numbered from 1 to n. Alice picks two balls at once. She places one of them in her left hand and the other in her right hand. Let N_1 denote the number of the ball that she holds in her left hand, and N_2 denote the corresponding number for her right hand. We define the random variable X to be equal to $N_1 - N_2$.

Alice performs this experiment many times. On some of the trials though, she crosses her hands, and thus her left hand appears to the right of her right hand. Alice likes to go by the

appearance, and thus on such instances, she computes X as $N_2 - N_1$.

Since both the balls had been picked simultaneously, her left and right hands are symmetric in this experiment, and crossing one over the other would make no difference to the expected value of X. Thus, if we denote the set of balls that she held with her left hand in multiple trials as L, and those that she held with her right hand as R, we can claim that on any individual trial i, exchanging L_i with R_i would make no difference to $\mathbb{E}[X]$, as L and R are symmetric, and it would appear as if she had originally picked L_i in her right hand and R_i in her left.

Summed up simply, symmetrization captures the notion that two training sets sampled independently from the same distribution are symmetric. Thus, exchanging a pair of points, one from the first set and one from the second set, should not make a difference to our analysis. It should proceed just as if the sets with exchanged points were the original first and second sets.

Now that we have all the tools that we need, we move on to our analysis, which would lead us to bound $\underset{S}{\mathbb{E}}[g_S]$.

4 Rademacher Complexity

In this section, we would try and bound the expected deviation of empirical risk from the l-risk of the supremum of a function class, and introduce the notion of Rademacher complexity. We would also look at tools that would help us in our analysis in the next section, when we finally derive a concentration bound for the case of linear regression.

4.1 Finding a bound for $\underset{S}{\mathbb{E}}[g_S]$

As was our objective, we want to limit the amount of expected difference between the empirical risk and the l-risk. We thus proceed as follows:

$$\begin{aligned}
\underset{S}{\mathbb{E}}[g_S] &= \underset{S}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\{er^l_S[f] - er^l_{\mathcal{D}}[f]\}\right] \\
&= \underset{S}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\left\{er^l_S[f] - \underset{\tilde{S}}{\mathbb{E}}\left[er^l_{\tilde{S}}[f]\right]\right\}\right] \\
&= \underset{S}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\left\{\underset{\tilde{S}}{\mathbb{E}}\left[er^l_S[f] - er^l_{\tilde{S}}[f]\right]\right\}\right] \\
&\leq \underset{S,\tilde{S}}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\{er^l_S[f] - er^l_{\tilde{S}}[f]\}\right] \\
&= \frac{1}{n}\underset{\{(x_i,y_i),(\tilde{x_i},\tilde{y_i})\}}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^{n} l(f(x_i),y_i) - l(f(\tilde{x_i}),\tilde{y_i})\right\}\right]
\end{aligned} \tag{2}$$

Here, the first of these equalities holds due to the definition of g_S, the second holds due to our first tool (expected empirical risk over $\tilde{S}$ is equal to the l-risk), and the third holds as S and $\tilde{S}$ are independently sampled, so $er^l_S[f] = \underset{\tilde{S}}{\mathbb{E}}\left[er^l_S[f]\right]$. Further, we use Jensen's inequality to upper bound the supremum of an expectation (using our third tool - the required supremum

here is a convex function).

Now, we are summing over the difference between the losses incurred at the i^{th} data points in S and $\tilde{S}$ for $i < n$. Notice here that S and $\tilde{S}$ are two training sets of the same size sampled independently from $\mathcal{D}$ and are thus symmetric. Thus, for a certain i, we can exchange the points that belong to S and $\tilde{S}$. Denoting (x_i, y_i) by z_i, we can push z_i to $\tilde{S}$ and $\tilde{z_i}$ to S, and our analysis must not change, by the argument of symmetrization.

We do this for $i = 1$, and our expression becomes:

$$\begin{aligned} &= \frac{1}{n} \mathop{\mathbb{E}}_{\{z_i, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \Big\{ \sum_{i=2}^{n} (l(z_i) - l(\tilde{z_i})) + l(\tilde{z_1}) - l(z_1) \Big\} \right] \\ &= \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_1, z_i, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \Big\{ \sum_{i=2}^{n} (l(z_i) - l(\tilde{z_i})) + \epsilon_1 (l(z_1) - l(\tilde{z_1})) \Big\} \right] \end{aligned} \tag{3}$$

where ϵ_1 is a Rademacher random variable that takes value 1 with probability 0.5 and takes -1 with probability 0.5, thus indicating that we admit an exchange with probability equal to that of no exchange, treating the two sets symmetrically. Proceeding in this manner for all i, we have,

$$\begin{aligned} &= \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_i, z_i, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \Big\{ \sum_{i=1}^{n} \epsilon_i (l(z_i) - l(\tilde{z_i})) \Big\} \right] \\ &\leq \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_i, z_i, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i l(z_i) + \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} (-\epsilon_i l(\tilde{z_i})) \right] \\ &= \frac{1}{n} \left(\mathop{\mathbb{E}}_{\{\epsilon_i, z_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i l(z_i) \right] + \mathop{\mathbb{E}}_{\{\epsilon_i, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} (-\epsilon_i l(\tilde{z_i})) \right] \right) \\ &= \frac{1}{n} \left(\mathop{\mathbb{E}}_{\{\epsilon_i, z_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i l(z_i) \right] + \mathop{\mathbb{E}}_{\{\tilde{\epsilon_i}, \tilde{z_i}\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \tilde{\epsilon_i} (\tilde{z_i}) \right] \right) \end{aligned} \tag{4}$$

The inequality stems from the fourth tool that we had stated: the supremum of the sum of two real number sequences is upper bounded by the sum of their suprema. Using linearity of expectation, we split the expectation operator over the two suprema. We also observe that $\{\epsilon_i\}$ can be chosen independently, and are symmetric about the origin, thus yielding the same distribution for $-\epsilon_i$ and ϵ_i, or equivalently, $\tilde{\epsilon_i}$.

Finally, we can observe that the two expectations are exactly the same with just a change of notation. This brings us to the crux of our analysis:

$$\begin{aligned} \mathop{\mathbb{E}}_{S} [g_S] &\leq \frac{2}{n} \mathop{\mathbb{E}}_{\{\epsilon_i, z_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i l(z_i) \right] \\ &= 2R_n(l \circ \mathcal{F}) \end{aligned} \tag{5}$$

where we define $R_n(l \circ \mathcal{F})$ to be the Rademacher complexity of a loss function class $l \circ \mathcal{F}$ as

$$R_n(l \circ \mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_i, x_i, y_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i l(f(x_i), y_i) \right] \tag{6}$$

where $\{(x_i, y_i)\}$ are drawn from the data distribution $\mathcal{D}$ independently for all $i \leq n$, and $\{\epsilon_i\}$ are independent Rademacher random variables. The loss function class $l \circ \mathcal{F}$ is induced by applying the loss function l to the function class $\mathcal{F}$.

Remark 9.2. Note that for every function f individually, $\mathop{\mathbb{E}}_{\{\epsilon_i, x_i, y_i\}} [\sum_{i=1}^{n} \epsilon_i l(f(x_i), y_i)]$ would be zero, but the expectation of the supremum may not be zero.

4.2 Ledoux-Talagrand Contraction Inequality

Rademacher complexity of a function class $\mathcal{F}$ is defined as:

$$R_n(\mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_i, x_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i) \right] \tag{7}$$

A low Rademacher complexity indicates that the training and test errors would be close for functions belonging to a function class $\mathcal{F}$. Note that it does not comment upon their magnitudes.

Remark 9.3. An empirical Rademacher complexity is defined when the training set S has already been sampled from a data distribution $\mathcal{D}$. It can be calculated as follows:

$$\hat{R}_n(\mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i) \Big| \{x_1, x_2, ..., x_n\} \right] \tag{8}$$

Remark 9.4. For the function class of all possible binary classifiers, $\mathcal{F} = \{-1, 1\}^{\mathcal{X}}$, there will always exist a function $f \in \mathcal{F}$ such that it will align exactly with $\{\epsilon_i\}$ at all i (turn 1 when ϵ_i is 1 for the i^{th} data point and vice versa), and would thus become the function at which the supremum attains its value and $R_n(\mathcal{F})$ becomes equal to 1, which is the highest value that it can take for classification. Thus, the Rademacher complexity of $\mathcal{F}$ captures the notion of its extreme expressiveness quite succinctly in this case.

Remark 9.5. It is also possible to choose $\{\epsilon_i\}$ as Gaussian random variables, since they are symmetric about the origin as well. This would lead us to develop the notion of Gaussian complexity. The interested reader may refer to Bartlett and Mendelson (2002) to gain further insight into Rademacher and Gaussian complexities.

In order to relate $R_n(\mathcal{F})$ and $R_n(l \circ \mathcal{F})$, we will need to introduce the notion of Lipschitzness of a loss function.

Definition 9.1. A loss function l is said to be L-Lipschitz ($L \in \mathbb{R}$) if, $\forall y, \hat{y_1}, \hat{y_2}$ in the domain,

$$|l(\hat{y_1}, y) - l(\hat{y_2}, y)| \leq L|\hat{y_1} - \hat{y_2}| \tag{9}$$

We use the Lipschitzness of our loss function crucially to move ahead with the rest of the analysis. We now introduce a contraction inequality that would help us relate $R_n(l \circ \mathcal{F})$ with $R_n(\mathcal{F})$.

Theorem 9.1. (Ledoux and Talagrand, 2013) If the function $l : (\mathbb{R} \times \mathbb{R}) \mapsto \mathbb{R}$ is L-Lipschitz, then

$$R_n(l \circ \mathcal{F}) \leq L R_n(\mathcal{F}) \tag{10}$$

The proof of this theorem is quite involved, and would not be covered in these notes. The reader is encouraged to look up the book cited above for more details.

5 Uniform Convergence for Linear Regression

We return to the setting of linear regression and try to bound the Rademacher complexity of our linear function class. This would help us bound the deviation of the training error from the test error with high probability using McDiarmid's inequality.

For linear regression, we take $\mathcal{W} = \mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\|_2 \leq R\}$. We also take our d-dimensional input space as $\mathcal{X} \subseteq \mathcal{B}_2(0, r) = \{x : \|x\|_2 \leq r\}$.

Now, the Rademacher complexity of function class $\mathcal{W}$ is bounded as follows:

$$R_n(\mathcal{W}) \leq \sup_{\{x_i\}} \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \langle w, x_i \rangle \right] \tag{11}$$

Here, we upper bound the expectation over the set S of data points $\{x_i\}$ drawn from $\mathcal{D}$ by the supremum of the quantity over all such sets.

$$\begin{aligned} &= \sup_{\{x_i\}} \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\sup_{w \in \mathcal{W}} \langle w, \frac{1}{n} \sum_{i=1}^{n} \epsilon_i x_i \rangle \right] \\ &\leq R \sup_{\{x_i\}} \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i x_i \right\|_2 \right] \end{aligned} \tag{12}$$

The inequality stems from the fact that for any vector v of the same dimensionality as w, $\sup_{w \in \mathcal{W}} \langle w, v\rangle = \sup_{\|w\|_2 \leq R} \langle w, v\rangle = R \|v\|_2$, by the Cauchy Schwartz inequality.

One problem that we run into here is dealing with the l_2 norm directly. It would be much easier to deal with the squared l_2 norm instead, so we invoke Jensen's inequality again to aid us in our time of need. z^2 is a convex function in variable z, and thus, we have, $(\mathbb{E}[z])^2 \leq \mathbb{E}[z^2]$.

Applying this to our cause, and taking square roots, we have,

$$\begin{aligned} R \sup_{\{x_i\}} \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i x_i \right\|_2 \right] &\leq R \sup_{\{x_i\}} \sqrt{\mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i x_i \right\|_2^2 \right]} \\ &= R \sup_{\{x_i\}} \sqrt{\mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\frac{1}{n^2} \Big(\sum_{i=1}^{n} \|x_i\|_2^2 + \sum_{i \neq j} \epsilon_i \epsilon_j \langle x_i, x_j \rangle \Big) \right]} \\ &= R \sup_{\{x_i\}} \sqrt{\frac{1}{n^2} \left(\sum_{i=1}^{n} \|x_i\|_2^2 + \mathop{\mathbb{E}}_{\{\epsilon_i\}} \left[\sum_{i \neq j} \epsilon_i \epsilon_j \langle x_i, x_j \rangle \right] \right)} \end{aligned} \tag{13}$$

Now,

$$\mathop{\mathbb{E}}_{\{\epsilon_i\}}\left[\sum_{i\neq j}\epsilon_i\epsilon_j\langle x_i,x_j\rangle\right]=\sum_{i\neq j}\mathop{\mathbb{E}}_{\epsilon_i,\epsilon_j}\left[\epsilon_i\epsilon_j\langle x_i,x_j\rangle\right] \tag{14}$$

This holds due to linearity of expectation. For a given pair i, j, ϵ_i and ϵ_j are independently sampled Rademacher random variables, and thus their product corresponds to another Rademacher random variable $\epsilon_{ij} = \epsilon_i\epsilon_j$ (as probability of the two having opposite signs is the same as that for the two having the same sign: $++, --, +-, -+$ are all equally likely). Thus, with probability 0.5, $\epsilon_i\epsilon_j\langle x_i, x_j\rangle = \langle x_i, x_j\rangle$, and with probability 0.5, $\epsilon_i\epsilon_j\langle x_i, x_j\rangle = -\langle x_i, x_j\rangle$. Thus the expected value of this product turns out to be zero for all i, j.

Therefore, the second term inside the square root goes to zero, and we have:

$$\begin{aligned}R_n(\mathcal{W}) &\leq \frac{R}{n}\sup_{\{x_i\}}\sqrt{\sum_{i=1}^{n}\|x_i\|_2^2}\\ &= \frac{Rr}{\sqrt{n}}\end{aligned} \tag{15}$$

This leads us to a very important result for linear regression with bounded weights - **the Rademacher complexity of the corresponding function class goes down as** $1/\sqrt{n}$.

Returning to the McDiarmid's inequality 1 and substituting for $\mathbb{E}[g_S]$, we have,

$$\mathop{\mathbb{P}}_{S}\left[\sup_{w\in\mathcal{W}}\{er_S^l[w]-er_{\mathcal{D}}^l[w]\}>\frac{2RrL}{\sqrt{n}}+\epsilon\right]\leq exp(\frac{-n\epsilon^2}{2B^2}) \tag{16}$$

If we let the error probability be δ, we have $\epsilon = B\sqrt{\frac{2log\frac{1}{\delta}}{n}}$,
and deviation $= \frac{2RrL}{\sqrt{n}} + \epsilon \leq 2(RrL + B)\sqrt{\frac{log\frac{1}{\delta}}{n}}$.

$$\mathop{\mathbb{P}}_{S}\left[\sup_{w\in\mathcal{W}}\{er_S^l[w]-er_{\mathcal{D}}^l[w]\}>2(RrL+B)\sqrt{\frac{log\frac{1}{\delta}}{n}}\right]\leq\delta \tag{17}$$

This means that with probability $1-\delta$, no function in the linear function class $\mathcal{W}$ is bad, i.e. has empirical risk and l-risk differing by a large quantity, if R, r and B are small. Further, this upper bound on deviation gets lower as the number of training points increases. Most notably, the deviation and the error probability are independent of the number of dimensions d of the input space, thus ensuring uniform convergence for very high dimensional input spaces as well.

6 Concluding Thoughts

In this lecture, we have seen for linear regression, how the expected difference between the empirical risk and the l-risk for all functions of a function class is upper bounded by a constant factor times its Rademacher complexity. Further, we have shown that this difference is large with a low probability, and the allowed "largeness" goes down with increasing number of samples taken for the empirical estimate. We have shown uniform convergence guarantees for linear

regression where the error probability is independent of the dimensionality of the input space, thus making it applicable for very large dimensional input spaces as well.

An intriguing observation that stems from these results is that even though the number of parameters in case of linear regression are d, we can get a high probability upper bound on the difference between the training and the test error with significantly less number of samples. For instance, when d is large, even $n < d$ samples may be enough to get a high probability bound on this difference. This is in stark contrast to the generic setting that we take in linear algebra, where we need d equations in d variables to be able to solve exactly for them. This apparent discrepancy emerges from the fact that we are constraining the d variables (in this case, such that $\|w\|_2 \leq R$) and thus are able to get nice solutions with far fewer equations than the number of variables.

References

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.