Lecture 5: Rademacher Complexity

Akshay Krishnamurthy
akshay@cs.umass.edu

September 25, 2017

1 Recap

Last time we introduced the VC dimension and saw one of the fundamental results in statistical learning theory. Recall that for a hypothesis space $\mathcal{H} : \mathcal{X} \to \{0, 1\}$, we say that $\mathcal{H}$ shatters a sample $C \subset \mathcal{X}$ if the $\mathcal{H}$ can realize all possible binary labelings of C. The VC dimension is the size of the largest set C that can be shattered by $\mathcal{H}$.

For binary classification with 0/1 loss, if the hypothesis space has VC dimension d then we can agnostically learn with

$$n_{\mathcal{H}}(\epsilon, \delta) \leq O\left(\frac{d\log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}\right).$$

samples. It turns out that VC-dimension also yields a sample complexity bound for the realizable case, and while we didn't prove it, the ideas behind the No Free Lunch theorem can be turned into sample complexity *lower bounds* for PAC-learning VC classes.

Theorem 1 (VC dimension lower bound)**.** *For every hypothesis class $\mathcal{H}$ with VCdim$(\mathcal{H}) = d < \infty$ and consider binary classification under 0/1 loss. Then the following sample complexity* lower bounds *apply*

1. *In the agnostic case:* $n_{\mathcal{H}}(\epsilon, \delta) \geq \Omega\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$
2. *In the realizable case:* $n_{\mathcal{H}}(\epsilon, \delta) \geq \Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$

Looking at the definition and the proof for the upper bound, VC dimension seems intimately tied to binary classification and 0/1 loss. A natural question is whether a similar notion exists for other learning problems and other loss functions. Today we discuss Rademacher complexity, which is precisely such a notion.

2 Uniform Convergence with Rademacher Complexity

To introduce Rademacher complexity, it is best to review the proof of the VC theorem. Recall that we wanted to bound the expected supremum

$$\mathbb{E}_{S\sim\mathcal{D}^n} \sup_h |R(h) - \hat{R}(h)|,$$

to show uniform convergence. This quantity is called an *empirical process*, and the study of such quantities is known as *empirical process theory*. Our strategy last time was to introduce a ghost sample and symmetrize:

$$\begin{aligned}
\mathbb{E}_S \sup_h R(h) - \hat{R}(h) &= \mathbb{E}_S \sup_h \mathbb{E}_{S'} \hat{R}'_S(h) - \hat{R}_S(h) \\
&\leq \mathbb{E}_{S,S'} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(x'_i) \neq y_i\} - \mathbf{1}\{h(x_i) \neq y_i\} \\
&= \mathbb{E}_{S,S'}\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i(\mathbf{1}\{h(x'_i) \neq y_i\} - \mathbf{1}\{h(x_i) \neq y_i\}).
\end{aligned}$$

This is the same proof we used last time (last time we worked on the absolute value, but you can just do both sides separately) and in the last line we have introduced the rademacher random variables $\epsilon_{1:n} \sim \text{Unif}(\{-1, 1\})$. This last expression is almost the definition of Rademacher complexity, for $\mathcal{H}$ and this is an intermediary term in the proof of the VC theorem. The last step is just to simplify slightly.

$$\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i(\mathbf{1}\{h(x_i') \neq y_i\} - \mathbf{1}\{h(x_i) \neq y_i\})| \leq \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \mathbf{1}\{h(x_i') \neq y_i'\} + \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\epsilon_i \mathbf{1}\{h(x_i) \neq y_i\}$$

And hence

$$\mathbb{E}_S \sup_h R(h) - \hat{R}(h) \leq 2\mathbb{E}_S\mathbb{E}_\epsilon \sup_h \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \mathbf{1}\{h(x_i) \neq y_i\}.$$

Notice that here we haven't really used the fact that the classifiers are binary, or that the loss function is the 0/1 loss. The exact same argument works for any loss function, and this reasoning leads to the definition of Rademacher complexity.

Definition 2 (Rademacher Complexity)**.** *For a set of vectors $A \subset \mathbb{R}^n$ the **rademacher complexity** is defined as $\mathcal{R}(A) = \frac{1}{n}\mathbb{E}_\epsilon \sup_{a\in A} \sum_{i=1}^n \epsilon_i a_i$.*

Remark 3. *The intuition is that $\mathcal{R}(A)$ captures how well A can fit random noise. To compute the rademacher complexity, we generate a random sign vector and find the vector $a \in A$ that has maximum inner product. Thus we fit random sign vectors with elements in A. If we can fit them well, then $\mathcal{R}(A)$ is big, but also intuitively you should expect to overfit in your learning problem.*

Rademacher complexity can explain uniform convergence in a very general setting, so we introduce a more general notation now. Let $\mathcal{Z}$ be an instance space, let $\mathcal{F}$ be a function class, and $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ be a loss function (think of $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ in the usual prediction set up). Then if D is a distribution over $\mathcal{Z}$, the natural empirical process of interest is

$$\mathbb{E}_{Z^n\sim\mathcal{D}^n} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i) - \mathbb{E}\ell(f, z).$$

The main theorem for Rademacher complexity relates this empirical process to the rademacher complexity of the class $\ell \cdot \mathcal{F} \cdot Z^n$, which is the set of vectors $\{(\ell(f, z_1), \ldots, \ell(f, z_n))\}_{f\in\mathcal{F}}$. We denote this as $\mathcal{R}(f \cdot \mathcal{F} \cdot Z^n)$.

Lemma 4 (Symmetrization with Rademacher Complexity)**.**

$$\mathbb{E}_{Z^n\sim\mathcal{D}^n} \sup_f R(f) - \hat{R}(f) \leq 2\mathbb{E}_{Z^n\sim\mathcal{D}^n}\mathcal{R}(\ell \cdot \mathcal{F} \cdot Z^n)$$

Proof. Essentially the proof is above. We reproduce the calculations briefly

$$\begin{aligned}\mathbb{E}_{Z^n\sim\mathcal{D}^n} \sup_f R(f) - \hat{R}(f) &\leq \mathbb{E}_{Z,Z'\sim\mathcal{D}^n} \sup_f \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i') - \ell(f, z_i) \\ &= \mathbb{E}_{Z,Z',\epsilon} \sup_f \frac{1}{n}\sum_{i=1}^{n} \epsilon_i(\ell(f, z_i') - \ell(f, z_i)) \leq 2\mathbb{E}_{Z^n}\mathcal{R}(\ell \cdot \mathcal{F} \cdot Z^n).\end{aligned}$$

□

Theorem 5. *Assume that $|\ell(f, z)| \leq c$ for all $f \in \mathcal{F}, z \in \mathcal{Z}$. Then with probability at least $1 - \delta$*

$$\sup_{f\in\mathcal{F}} R(f) - \hat{R}(f) \leq 2\mathbb{E}_{Z'}\mathcal{R}(\ell \cdot \mathcal{F} \cdot Z') + c\sqrt{2\log(2/\delta)/n} \tag{1}$$

$$\sup_{f\in\mathcal{F}} R(f) - \hat{R}(f) \leq 2\mathcal{R}(\ell \cdot \mathcal{F} \cdot Z') + 4c\sqrt{2\log(4/\delta)/n} \tag{2}$$

Proof. The proof of the first inequality requires applying Lemma 4 and McDiarmid's inequality. For the second, we additionally apply McDiarmid's inequality to the Rademacher complexity, which leads to the *empirical rademacher complexity*. (Exercise: complete the proof on your own). □

Data-dependent bounds. As another remark, the bound in Eq. (2) is a *data dependent bound* which is quite useful and can be much smaller in practice than the expected bound in Eq. (1), or even the worst case VC bound we saw last time. If you have a benign problem, then the expected rademacher complexity can be much better than the worst-case VC-dimension, and if you have a favorable sample, it can be even smaller! In this way Rademacher complexity can lead to much tighter sample complexity bounds. Obtaining such data-dependent bounds in other settings is an active area of research.

Implications for learning. Since we know that uniform convergence is sufficient for agnostic learnability, we see that whenever the Rademacher complexity is well-behaved, learning is possible. Thus we need to develop a set of tools for working with and understanding Rademacher complexities.

3 Rademacher Calculus

Rademacher complexity is a fairly easy quantity to work with and satisfies a number of useful properties that support various operations. Here we will prove some of them

Lemma 6 (Rademacher bound for finite classes (Massart Lemma)). *Let $A = \{a^{(1)}, \ldots, a^{(N)}\}$ be a finite set of vectors in $\mathbb{R}^n$ and define $\bar{a} = \frac{1}{N}\sum_{i=1}^n a^{(i)}$ to be the average. Then,*

$$\mathcal{R}(A) \leq \max_a \|a - \bar{a}\| \frac{\sqrt{2\log(N)}}{n}$$

This lemma helps us recover uniform convergence for VC-classes, since for a VC-class on a sample Z^n the number of possible vectors is at most the growth function $\tau_n(\mathcal{H})$. In particular for 0/1 loss, the vectors A we use are the vectors $\{(\mathbf{1}\{h(x_1) \neq y_1\}, \ldots, \mathbf{1}\{h(x_n) \neq y_n\})\}_{h\in\mathcal{H}_C}$ where $\mathcal{H}_C$ is the restriction. These vectors have norm at most $\sqrt{n}$ so this recovers the $\sqrt{\log(\tau_n(\mathcal{H}))/n}$ rate that we saw in the VC theorem.

As you might suspect the proof uses the same log-sum-exp method that we saw in the VC theorem proof.

Proof. Assume that $\bar{a} = 0$, which is without loss of generality since the rademacher complexity is translation invariant.

$$\begin{aligned}
n\mathcal{R}(A) &= \mathbb{E}_\epsilon \max_{a\in A}\langle\epsilon, a\rangle = \frac{1}{\lambda}\mathbb{E}\log\exp\max_{a\in A}\lambda\langle\epsilon, a\rangle \\
&\leq \frac{1}{\lambda}\log\mathbb{E}\exp(\max_a \lambda\langle\epsilon, a\rangle) \\
&\leq \frac{1}{\lambda}\log\sum_a\prod_{i=1}^n \mathbb{E}\exp(\lambda\epsilon_i, a_i) \\
&\leq \frac{1}{\lambda}\log\sum_a\prod_{i=1}^n \exp(\lambda^2 a_i^2/2) = \frac{1}{\lambda}\log|A| + \lambda\max_a\|a\|_2^2/2
\end{aligned}$$

Here we first use Jensen's inequality, then replace max with sum, then apply the MGF bound for rademacher random variables we saw previously. The final step is to optimize for λ, which by taking derivative and setting to zero, reveals that we should set $\lambda = \sqrt{2\log(|A|)/\max_a\|a\|_2^2}$ and proves the lemma. □

Another related property that is quite useful is that Rademacher complexity of a convex hull of a point set is the same as the rademacher complexity of the point set itself.

Lemma 7 (Rademacher bound for convex hulls). *Let $A \subset \mathbb{R}^n$ and define*

$$conv(A) \triangleq \{\sum_{j=1}^n \alpha_j a^{(j)} : N \in \mathbb{N}, a^{(j)} \in A, \alpha_j \geq 0, \sum_j \alpha_j = 1\},$$

to be the set of all distributions over A. Then $\mathcal{R}(A) = \mathcal{R}(conv(A))$.

This lemma is extremely useful since it means that learning over distributions of hypotheses is statistically not any harder than learning over individual hypotheses. Indeed we will see this happen later in the course when we discuss online learning. This property is also useful in convex relaxations, which lead to computationally tractable solutions to problems in sparse recovery.

Example 1. *In sparse linear regression, we want to find a linear predictor with at most s-nonzero coordinates. Here the loss function is the square loss $\ell(y, y') = (y - y')^2/2$ and the function class is defined in terms of a ball $\mathcal{W}_0 = \{w \in \mathbb{R}^d, \|w\|_2^2 \le 1, \|w\|_0 \le s\}$ as $\mathcal{F}_0 = \{f_w(x) = \langle w, x\rangle, w \in \mathcal{W}_0\}$. The ERM problem here is*

$$\operatorname*{argmin}_w \frac{1}{2n}\sum_{i=1}^n (\langle w, x_i\rangle - y_i)^2 \ \ s.t.\ \|w\|_2^2 \le 1, \|w\|_0 \le s$$

which unfortunately is NP-hard in the worst case. However, if we work over the ℓ_1 ball $\mathcal{W}_1 = \{w \mid \|w\|_2^2 = 1, \|w\|_1 = \sum_i |w_i| \le \sqrt{s}\}$ then the optimization problem is convex, but moreover you can prove that

$$conv(\mathcal{W}_0) \subset \mathcal{W}_1 \subset 2conv(\mathcal{W}_0),$$

so $\mathcal{R}(\mathcal{W}_1) \le 2\mathcal{R}(\mathcal{W}_0)$. Thus, at least statistically, working with the ℓ_1 ball does not substantially hurt you.

Proof of Lemma 7. The key idea here is that maximizing a linear function over the convex hull always yields a vertex, which is by definition a point in the original set A. Or more precisely, if v is a vector, then

$$\sup_{\alpha \ge 0, \sum \alpha_j = 1} \sum_j \alpha_j v_j = \max_j v_j.$$

This can be seen by the fact that on the left you want to put all the mass of α onto the coordinate with largest magnitude.

$$n\mathcal{R}(\text{conv}(A)) = \mathbb{E}_\epsilon \sup_\alpha \sup_{a^{(1)},\ldots,a^{(N)}} \sum_{i=1}^n \epsilon_i \sum_{j=1}^N \alpha_j a_i^{(j)} = \mathbb{E}_\epsilon \sup_\alpha \sum_{j=1}^N \alpha_j \sup_{a^{(j)}} \sum_i \epsilon_i a_i^{(j)} = \mathbb{E}_\epsilon \sup_a \sum_{i=1}^n \epsilon_i a_i = n\mathcal{R}(A) \qquad \square$$

The final lemma helps with composing loss functions

Lemma 8 (Contraction Lemma)**.** *For each $i \in [m]$ let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function (that is $|\phi_i(x) - \phi_i(y)| \le \rho|x - y|$ for all x, y). For $a \in \mathbb{R}^n$ let $\phi(a) = (\phi_1(a_1)\ldots, \phi_n(a_n))$ and for a set A, let $\phi \cdot A = \{\phi(a) \mid a \in A\}$. Then $\mathcal{R}(\phi \cdot A) \le \rho\mathcal{R}(A)$.*

Proof. The key idea in the proof here is to explicitly write out the expectation for one random variable and then use the Lipschitz property on that variable.

$$\begin{aligned}
n\mathcal{R}(\phi \cdot A) &= \mathbb{E}_\epsilon \sup_a \sum_{i=1}^n \epsilon_i \phi_i(a_i) \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_a \left\{\phi_1(a_1) + \sum_{i=2}^n \epsilon_i\phi_i(a_i)\right\} + \sup_a \left\{-\phi_1(a_1) + \sum_{i=2}^n \epsilon_i\phi_i(a_i)\right\} \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{a,a'} \phi_1(a_1) - \phi_1(a_1') + \sum_{i=2}^n \epsilon_i\phi_i(a_i) + \sum_{i=2}^n \epsilon_i\phi_i(a_i') \\
&\le \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{a,a'} \rho|a_1 - a_1'| + \sum_{i=2}^n \epsilon_i\phi_i(a_i) + \sum_{i=2}^n \epsilon_i\phi_i(a_i') \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{1:n}} \sup_{a,a'} \rho\epsilon_1(a_1 - a_1') + \sum_{i=2}^n \epsilon_i\phi_i(a_i) + \sum_{i=2}^n \epsilon_i\phi_i(a_i') \\
&= \mathbb{E}_{\epsilon_{1:n}} \sup_a \rho\epsilon_1 a_1 + \sum_{i=2}^n \epsilon_i\phi_i(a_i).
\end{aligned}$$

The inequality uses the Lipshitz property. The absolute value can be removed since the optimization for a, a' is over the same set and the remainder of the expression does not change if we swap the names. $\square$

4 Why Rademacher?

Why is Rademacher complexity a better notion than VC-dimension?

1. The answer that motivated our discussion is that we want a theory that applies to other prediction problems and also other loss functions. Rademacher complexity can be used to derive sample complexity bounds for regression, classification with other loss functions like hinge loss of logistic loss, and even unsupervised learning problems.

2. Another reason we saw briefly is that we can get data-dependent bounds Eq. (2). This can help since we can often measure the empirical Rademacher complexity, so we can know if we are overfitting or not.

3. Moreover, the Rademacher calculus we have seen means that rademacher complexity bounds can be easily applied to a number of problems.

Excess risk bounds for linear regresssion. Consider linear regression with the square loss $\ell(x, y) = 1/2(x-y)^2$. So $\mathcal{W}_2 = \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}$ and $\mathcal{F}_w = \{f_w(x) = \langle w, x\rangle, w \in \mathcal{W}_2\}$. Let $\mathcal{Z} = B_2(1) \times [-1, 1]$ be the example space. Then

$$\mathcal{R}(\ell \circ \mathcal{F} \circ Z^n) \leq 2/\sqrt{n}.$$

The calculation is as follows:

$$n\mathcal{R}(\ell \circ \mathcal{F} \circ Z^n) \leq 2n\mathcal{R}(\mathcal{F} \circ Z^n) = 2\mathbb{E}_\epsilon \sup_w \sum_{i=1}^n \epsilon_i \langle w, x_i\rangle = 2\mathbb{E}_\epsilon \sup_w \langle w, \sum_{i=1}^n \epsilon_i x_i\rangle = 2\mathbb{E}_\epsilon \| \sum_i \epsilon_i x_i\|$$

$$\leq 2\left(\mathbb{E}\| \sum_i \epsilon_i x_i\|_2^2\right)^{1/2} = 2\left(\sum_{i,j} \langle x_i, x_j\rangle \mathbb{E}\epsilon_i\epsilon_j\right)^{1/2} \leq 2\sqrt{n}.$$

In the first step, we use that ℓ_y is 2-Lipschitz, since we can write $(f(x) - y)^2 - (g(x) - y)^2 \leq (f(x) - g(x))(f(x) + g(x) - 2y) \leq 4|f(x) - g(x)|$ owing to the boundedness properties. The square root arises through an application of Jensen's inequality.

Thus we managed to prove an excess risk bound for linear regression, which is what we introduced in the first lecture. However, note that there we promised an $O(d/n)$ sample complexity bound, while here we are showing a $O(1/\sqrt{n})$ bound. There are two main differences. First this bound applies in an agnostic setting, so the linear model need not be true, and this leads to a $O(1/\sqrt{n})$ rate instead of the better $O(1/n)$ rate. Second, this bound operates over a function class with bounded ℓ_2 norm, rather than the unconstrained class we used in Lecture 1. This is closer to ridge regression, and can have substantial impact on the sample complexity.