CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 20**

Operations with Rademacher Averages

Lecturer: Peter Bartlett *Scribe: Leon Barrett*

1 Introduction

Previously, we related the empirical risk, $\widehat{L}$, to the risk, L in a uniform way so as to reason about convergence. We were able to put such bounds on the risk by way of Rademacher averages, R_n, which we could bound for special cases. For example, in the case where $F \subseteq \{\pm 1\}^{\mathcal{X}}$ is finite, the bound is

$$R_n(F) \leq \sqrt{\frac{2 \log |F|}{n}}$$

As another example, if we can restrict the growth function, then the bound is

$$R_n(F) \leq \sqrt{\frac{\log 2\Pi_F(n)}{n}}$$

Recall that slow growth of Π, the growth function, is equivalent to a low VC dimension, because the growth function increases exponentially up to the VC dimension, after which it grows polynomially.

If we define Regret $= L(\hat{f}) - L^*$ then it suffices to show that $\widehat{L}(f) \to L(f)$ *uniformly*. Uniform convergence depends upon the complexity of the function class F. In this lecture, we study techniques to get bounds for Rademacher averages of general function classes.

2 Bounds on regret

Theorem 2.1. With probability at least $1-\delta$, the risk $L(\widehat{f})$ satisfies the following:

$$\begin{aligned} L(\widehat{f}) &\leq \inf_{f \in F} L(f) + \sqrt{\frac{2 \log \Pi_F(n)}{n}} + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} \\ &\leq \inf_{f \in F} L(f) + \sqrt{\frac{2d \log(en/d)}{n}} + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} \end{aligned}$$

where $d_{VC}(F) = d$ and $\widehat{f}$ is the function in $F \subseteq \{\pm 1\}^{\mathcal{X}}$ that minimizes $\widehat{L}$. The second term, $\sqrt{\frac{2d \log(en/d)}{n}}$, represents a penalty for the complexity of the class F.

Today, we will look at ways of getting bounds for classes by examining how composition and other operations change the Rademacher averages.

2.1 Lower bounds on loss

First, let us examine a lower bound on the expected loss of a function class. We will do that with the following converse of Theorem 2.1:

Theorem 2.2. Fix class F with $d_{VC}(F) = d$ and $n \geq 1$. For any mapping

$$f_n : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y}$$

where $\mathcal{Y} = \{\pm 1\}$, there is a probability distribution P on $\mathcal{X} \times \mathcal{Y}$ such that with probability $\geq \frac{1}{200}$ (some constant), the expected loss is at least the best in class plus a margin. That is,

$$\mathbb{E}\left[l(Y, f_n(X_1, Y_1, \ldots, X_n, Y_n; X)) | (X_i, Y_i)_1^n\right] \geq \inf_{f \in F} \mathbb{E}l(Y, f(X)) + c \min(\sqrt{\frac{d}{n}}, 1)$$

where c is a universal constant.

Proof. The proof idea, which is all we give here, uses the probabilistic method. We choose P randomly from a set concentrated on a set shattered by F, with

$$P(Y = 1 | X = x) \in \frac{1}{2} \pm \epsilon$$

That is, at x, it is very hard to distinguish whether it is more likely that $Y = 1$ or $Y = 0$. Getting close to the optimum requires having seen a significant number of examples from each position of the set. So at an x that you have not seen, learning depends on how fast you can get the estimated probability to the correct side of $\frac{1}{2}$. As d gets big, there are more areas to visit, so this gets harder. □

The point is that while any particular distribution may have $L(\widehat{f})$ drop rapidly, there remains a distribution such that the expected loss drops slowly. See, for instance, Figure 2.1. Note that the "bad" distribution

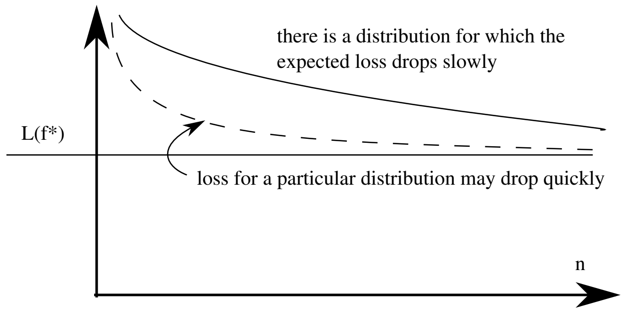

Figure 1: $L(\widehat{f})$ may drop rapidly for some distribution, but there is a distribution such that the expected loss drops slowly.

changes with the sample size. There is a paper by Dale Schuurmans[1] addressing the case when we require the distribution to remain fixed with the sample size. He essentially deals with classes that have $d_{VC} = 1$. There are also types of dimensions other than d_{VC}, such as ones useful for regression, that give similar results.

[1]Schuurmans, D. Characterizing rational versus exponential learning curves. In *Proceedings of the Second European Conference on Computational Learning Theory (EuroCOLT-95)*, pp. 272–286. (1995)

3 Working with Rademacher averages

How can we use knowledge about Rademacher averages and compositions of functions to learn something about the risk for a class? Ideally, we'd like to be able to start with simple classes and build up more complex classes, while keeping the regret bounds that the Rademacher averages give us. To that end, we will look at some properties of Rademacher averages ($R_n(F)$). Recall that

$$R_n(F) = \mathbb{E}\widehat{R}_n(F)$$

and

$$\widehat{R}_n(F) = \mathbb{E}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\Big| X_1,\ldots,X_n\right]$$

Properties of Rademacher averages

1. $$F \subseteq G \Rightarrow R_n(F) \le R_n(G) \tag{1}$$

2. $$R_n(cF) = |c|R_n(F)$$

   where

   $$cF = \{cf : f \in F\} \tag{2}$$

   This can be seen by examining the formula for Rademacher averages.

   $$\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}(c\epsilon_i)f(X_i)$$

   Then $c\epsilon_i$ has the same distribution as $|c|\epsilon_i$, and the $|c|$ term can be pulled out:

   $$|c|\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)$$

   From there, it is easy to derive Property 2.

3. Let

   $$F + g = \{f + g : f \in F\}$$

   Then

   $$R_n(F+g) = R_n(F) \tag{3}$$

   because the g term vanishes as follows.

   $$\begin{aligned}
\mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f+g)(X_i) &= \mathbb{E}\left[\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right] \\
&= \mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) + \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i) \\
&= \mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) + 0
\end{aligned}$$

4. Let co(F) represent the convex hull of F. Specifically,

$$\mathrm{co}(F) = \left\{ \sum_{i=1}^{m} \alpha_i f_i : \alpha_i \geq 0,\ \sum_{i=1}^{m} \alpha_i = 1,\ f_i \in F \right\}$$

Also, let

$$-F = \{-f : f \in F\}$$

Then

$$R_n(F \cup -F) = R_n(\mathrm{co}(F \cup -F)) \tag{4}$$

To see this, note that Property 1 implies

$$R_n(F \cup -F) \leq R_n(\mathrm{co}(F \cup -F))\ .$$

But,

$$R_n(\mathrm{co}(F \cup -F)) = \mathbb{E} \sup_{m} \sup_{f_1 \ldots f_m \in F} \sup_{||\alpha||_1 = 1} \left(\sum_{j=1}^{m} \alpha_j \underbrace{\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(X_i)}_{s_j} \right)$$

Then

$$\begin{aligned} R_n(\mathrm{co}(F \cup -F)) &= \mathbb{E} \sup_{m} \sup_{f_1 \ldots f_m \in F} \sup_{||\alpha||_1 = 1} \left(\sum_{j=1}^{m} \alpha_j s_j \right) \\ &= \mathbb{E} \sup_{m} \sup_{f_1 \ldots f_m \in F} \sup_{j \in \{1, \ldots, m\}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(X_i) \right| \end{aligned}$$

This is because

$$\sup_{||\alpha||_1 = 1} \alpha' s = \max_{j} |s_j|\ .$$

The left hand side is the maximum of a linear function over the L_1 ball and the maximum must be attained at a corner of that L_1 ball. So, we finally get

$$\begin{aligned} R_n(\mathrm{co}(F \cup -F)) &= \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \\ &= \mathbb{E} \sup_{f \in F \cup -F} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right) \\ &= R_n(F \cup -F) \end{aligned}$$

5. For $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfying

$$|\phi_i(a) - \phi_i(b)| \leq B|a - b|$$

for some constant B (the "Lipschitz constant"), we have

$$\mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(f(X_i)) \leq B \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \tag{5}$$

This is the Ledoux-Talagrand contraction inequality[2].

[2] Refer to Theorem 4.12 in Talagrand's book, *Probability in Banach Spaces*. The book is available online at his homepage `http://www.proba.jussieu.fr/users/talagran/`

3.1 Example

Take the following function class and bounded loss function:

$$F \in \mathbb{R}^{\mathcal{X}}$$

$$l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$$

Suppose, $\forall Y$, that the difference in l is bounded: $|l(Y, a) - l(Y, b)| \leq B|a - b|$. For example, it could be the truncated hinge loss, as seen in Figure 2. Then, by Property 5, with $\phi_i(f(X_i)) = l(Y_i, f(X_i))$, we have

$$R_n(l_F) \leq BR_n(F) \ .$$

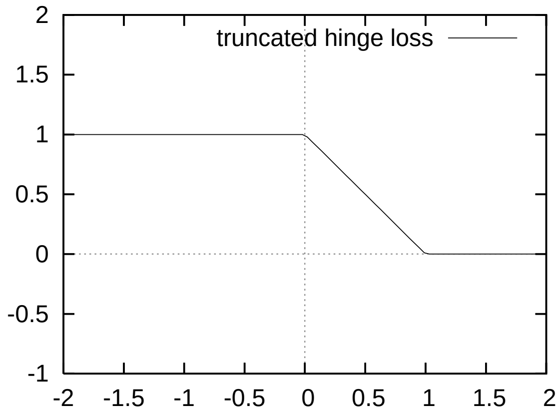

Figure 2: The truncated hinge loss