CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 17**

Rademacher Averages and Growth Functions

Lecturer: Peter Bartlett *Scribe: Kurt Miller*

In previous lectures, we showed that with probability $\geq 1-\delta$,

$$R(\hat{f}) \leq \inf_{f \in F} R(f) + 2R_n(\ell_F) + c\sqrt{\frac{\log(1/\delta)}{n}}, \tag{1}$$

where

$$\hat{f} = \arg\min_{f \in F} \hat{R}(f)$$

and $R_n(F)$ is the Rademacher average

$$R_n(F) = \mathbb{E}\left[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)\right].$$

We have also proved various properties of Rademacher averages. The properties relevant here are:

1. When scaling F by a constant factor c,
$$R_n(cF) = |c| R_n(F).$$
2. The Ledoux-Talagrand contraction inequality: for ℓ-Lipschitz functions ϕ
$$R_n(\phi \circ F) \leq \ell R_n(F).$$

Example 1

Let $F \subseteq \mathbb{R}^{\mathcal{X}}$, $\phi : \mathbb{R} \mapsto [0,1]$ be 1-Lipschitz, and consider $\ell_F = \{(x,y) \mapsto \phi(yf(x)) : f \in F\}$. For example, ϕ could be the truncated hinge loss $\phi(\alpha) = \min(1, \max(1-\alpha, 0))$. Then, using the contraction mapping inequality, with probability $\geq 1-\delta$

$$\begin{aligned} \mathbb{E}\phi(Yf(X)) - \hat{\mathbb{E}}(\phi(Yf(X))) &\leq 2R_n(\ell_F) + \sqrt{\frac{\log(1/\delta)}{2n}} \\ &\leq 2R_n(F) + \sqrt{\frac{\log(1/\delta)}{2n}}. \end{aligned}$$

This can also be applied to the truncated exponential loss $\phi(\alpha) = \min(1, e^{-\alpha})$.

Example 2

Let $G \subseteq \{\pm 1\}^{\mathcal{X}}$ with $d_{\mathrm{VC}}(G) < \infty$. Let $F = \lambda \mathrm{co}(G) = \{x \mapsto \sum \alpha_i g_i(x) : g_i \in G, \sum \alpha_i = \lambda, \alpha_i \geq 0\}$, i.e. F is the convex hull of G scaled by λ, a constant > 0. Then, for a constant c which includes the Lipschitz constant of our loss ϕ,

$$\begin{aligned} R_n(F) &= \lambda R_n(\mathrm{co}(G)) \\ &= \lambda R_n(G) \\ &= c\lambda\sqrt{\frac{d_{\mathrm{VC}}(G)}{n}}. \end{aligned}$$

So therefore

$$R_\phi(\hat{(}f)) \leq \inf_{f\in F} R_\phi(f) + c\lambda\sqrt{\frac{d_{\mathrm{VC}}(G)+\log(1/\delta)}{n}}.$$

As λ increases, the optimal risk $\inf_{f\in F} R_\phi(f)$ decreases, but the second term increases, so there is a tradeoff when choosing λ.

1 Rademacher Averages of Kernel Classes

Let F be a kernel class. We have previously seen the optimization

$$\text{minimize}_f \quad c\hat{\mathbb{E}}\phi(Yf(X)) + \|f\|_{\mathcal{H}}$$

for RKHS $\mathcal{H}$. For appropriate settings of Lagrangian multipliers, this is equivalent to

$$\text{minimize}_{f:\|f\|_{\mathcal{H}}\leq B} \quad c\hat{\mathbb{E}}\phi(Yf(X)).$$

We therefore wish to look at the Rademacher average $R_n(F_B)$ of $F_B = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq B\}$. Let K be the kernel matrix for $X_1^n = \{x_1, \cdots, x_n\}$ using the reproducing kernel for $\mathcal{H}$ so that $K_{ij} = k(x_i, x_j)$.

Theorem 1.1.

$$\begin{aligned}
\hat{R}_n(F_B) &\equiv \mathbb{E}\left[\sup_{f\in F_B}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\middle| X_1^n\right] \\
&\leq \frac{B}{n}\sqrt{\mathrm{trace}(K)} \\
&= \frac{B}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n k(x_i,x_i)}
\end{aligned}$$

Also if $\{\lambda_j\}$ are the eigenvalues of $T_k : f \mapsto \int k(\cdot, x)f(x)dP(x)$, then

$$R_n(F_B) \leq B\sqrt{\sum_{i=1}^\infty \lambda_i/n}$$

Proof. Using properties of the reproducing kernel and linearity,

$$\begin{aligned}
\sup_{f\in F_B}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) &= \sup_{f\in F_B}\frac{1}{n}\sum_{i=1}^n \epsilon_i\langle k(x_i,\cdot), f\rangle \\
&= \sup_{f:\|f\|_{\mathcal{H}}\leq B}\left\langle \frac{1}{n}\sum_{i=1}^n \epsilon_i k(x_i,\cdot), f\right\rangle \\
&= B\frac{\|\frac{1}{n}\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\|^2}{\|\frac{1}{n}\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\|} \\
&= B\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\right\| \\
&= B\sqrt{\frac{1}{n^2}\sum_{i,j}^n \epsilon_i\epsilon_j k(x_i,x_j)}
\end{aligned}$$

Therefore by an application of Jensen's inequality and using the fact that ϵ_i are i.i.d. with $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = 1$,

$$\begin{aligned}
\hat{R}_n(F_B) &= \mathbb{E}\left[\frac{B}{n}\sqrt{\left|\sum_{i,j}^{n}\epsilon_i\epsilon_j k(x_i,x_j)\right|}\,\middle|\, X_1^n\right] \\
&\leq \frac{B}{n}\sqrt{\mathbb{E}\left[\left|\sum_{i,j}^{n}\epsilon_i\epsilon_j k(x_i,x_j)\right|\,\middle|\, X_1^n\right]} \\
&= \frac{B}{n}\sqrt{\sum_{i}^{n}k(x_i,x_i)} \\
&= \frac{B}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i}^{n}k(x_i,x_i)}
\end{aligned}$$

where the last line has been rewritten to emphasize the fact that this is function of the average of $k(x_i, x_i)$.

Furthermore, since $R_n(F_B) = \mathbb{E}\hat{R}_n(F_B)$, then using the above result along with Jensen's inequality again, we get that

$$\begin{aligned}
R_n(F_B) &= \mathbb{E}\hat{R}_n(F_B) \\
&\leq \mathbb{E}\frac{B}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i}^{n}k(x_i,x_i)} \\
&\leq \frac{B}{\sqrt{n}}\sqrt{\mathbb{E}\frac{1}{n}\sum_{i}^{n}k(x_i,x_i)} \\
&= \frac{B}{\sqrt{n}}\sqrt{\mathbb{E}k(x,x)}.
\end{aligned}$$

Now using the fact that $k(x,y) = \sum_{i=1}^{\infty}\lambda_i\psi_i(x)\psi_i(y)$ for an orthonormal eigenbasis ψ_i, then we get

$$\begin{aligned}
R_n(F_B) &\leq \frac{B}{\sqrt{n}}\sqrt{\mathbb{E}k(x,x)} \\
&\leq \frac{B}{\sqrt{n}}\sqrt{\sum_{i=1}^{\infty}\lambda_i}
\end{aligned}$$

as desired.

Therefore

$$\begin{aligned}
\mathbb{E}\phi(Yf(X)) &\leq \hat{\mathbb{E}}\phi(Yf(X)) + 2R_n(\ell_F) + \sqrt{\log(1/\delta)/2n} \\
&\leq \hat{\mathbb{E}}\phi(Yf(X)) + \frac{cB}{\sqrt{n}}\sqrt{\sum_{i=1}^{\infty}\lambda_i} + \sqrt{\log(1/\delta)/2n}
\end{aligned}$$

where c is a constant that includes the Lipschitz constant of ϕ used in our loss function. □

2 Growth Functions

We have defined the growth function $\Pi_F(n)$ to be

$$\Pi_F(n) = \max\left\{|F_{|X_1^n}| : \{x_1, \cdots, x_n\} \subseteq \mathcal{X}\right\}$$

and have shown that

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

for $F \subseteq \{\pm 1\}^{\mathcal{X}}$. In other words, $\Pi_F(n)$ is the maximum number of distinct labelings that functions $f \in F$ can assign to any set of n points in $\mathcal{X}$. Therefore, by definition, $d_{\text{VC}} = \max\{n : \Pi_F(n) = 2^n\}$.

Motivation: We wish to compute the growth function for parameterized binary functions

$$F = \{x \mapsto f(x, \theta) : \theta \in \Theta\}$$

where $\Theta \subseteq \mathbb{R}$. If we can bound the growth function, then we can bound the risk of $\hat{f}$ in equation (1). For now, we will focus on the special case of linear threshold functions

$$F = \{x \mapsto \text{sign}(w'x - \theta) : w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$$

where (for concreteness), we let

$$\text{sign}(\alpha) = \begin{cases} 1 & \alpha \geq 0 \\ -1 & \alpha < 0 \end{cases}.$$

Theorem 2.1. For the class F of linear threshold functions

$$\Pi_F(n) = 2\sum_{i=0}^{d} \binom{n-1}{i}.$$

Proof sketch. We provide only a sketch of the proof here[1].

Start by fixing a set of points $S \subseteq \mathbb{R}^d$ where $|S| = n$. The idea of the proof is to divide the parameter space of $(w, \theta) = \mathbb{R}^{d+1}$ into "decision equivalence classes." We will show that there are finitely many such equivalence classes and that they can be counted by a geometric argument originally given by Schaffli in 1851.

1. Assume the points in S are in "general position," i.e. all subsets

$$\left\{\binom{x_1}{1}, \binom{x_2}{1}, \cdots, \binom{x_n}{1}\right\}$$

of size $\leq d+1$ are linearly independent. This implies that no three points are in a line, no four are in a plane, etc. If this is not true, then note that a small random perturbation will put the points in general position.

2. For each $x_i \in S$, define the hyperplane

$$P_i = \left\{(w, \theta) \in \mathbb{R}^{d+1} : w'x_i + \theta = 0\right\}.$$

[1]The full version of this proof can be found in *Neural Network Learning: Theoretical Foundations* by M. Anthony and P. Bartlett, pages 30–35. This proof emphasizes the link to the parameterized classes that we consider in the next lecture.

In order for (w, θ) and (w', θ') to label x_i differently, they must lie on opposite sides of P_i (assuming neither is on P_i). So define the set of connected components (CC) in $\mathbb{R}^{d+1}$ when split by all P_i to be

$$\begin{aligned} \left|F_{|S}\right| &= \mathrm{CC}(\mathbb{R}^{d+1} \setminus \cup_{i=1}^{n} P_i) \\ &\equiv C(n, d+1). \end{aligned}$$

Note that $C(n, d+1)$ does not depend on the actual choice of S. It only depends on the number of points n and the dimensionality of the space d that they lie in. The key here is that all (w, θ) in the same connected component label each point identically. For each (w, θ) and (w', θ') in different connected components, they label at least one point differently. Therefore, the number of connected components directly corresponds to the number of different labelings of S that can be achieved by F.

3. We first note that $C(1, d) = 2 \ \ \forall d$. This is because in any dimension with 1 point, P_1 will always split $\mathbb{R}^d$ into two.

4. Next, we show $C(n+1, d) = C(n, d) + C(n, d-1)$. This is because $C(n, d)$ corresponds to how many connected components there are with only n points. When we add the $n+1^{\text{th}}$ point, all these components still exist, but some of them are broken in two. This means that $C(n+1, d) = C(n, d)+$ the number of components in $C(n, d)$ that were split by P_{n+1}. This additional value is equal to the number of connected components of $P_{n+1} \setminus \cup_{i=1}^{n} P_i$, which is equal to $c(n, d-1)$.

5. By induction, this shows that $C(n, d) = 2 \sum_{k=0}^{d-1} \binom{n-1}{k}$, which proves the desired result.

□

This gives us the growth function for linear threshold functions. In the next lecture, we will give a similar result for more general parameterized binary classes

$$F = \{x \mapsto f(x, \theta) : \theta \in \Theta\}.$$