CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Statistical Learning Theory—Symmetrization Lemma

Lecturer: Michael I. Jordan *Scribes: Alice X. Zheng*

In the last few lectures, we proved bounds for the difference between empirical and real risks for a finite set of functions.[1] Our goal now is to prove similar results for classes containing an infinite number of functions. We will see that in the more general bounds, the number of functions N is replaced by H^Λ_{ann}, which is a better measure of complexity of an infinite set of functions. More specifically, we define $N^\Lambda(z_1, \ldots z_\ell)$ as the number of different partitions of $z_1, \ldots z_\ell$ which can be represented by functions in Λ. Recall the definitions for entropy and annealed entropy of a class of functions $\Lambda = \{\alpha\}$.

$$\begin{aligned} \text{Entropy}: \quad H^\Lambda(z_1, \ldots z_\ell) &= \ln N^\Lambda(z_1, \ldots z_\ell) && (1) \\ \text{Annealed entropy}: \quad H^\Lambda_{ann}(\ell) &= \ln EN^\Lambda(z_1, \ldots z_\ell) && (2) \end{aligned}$$

Our goal is to prove the following theorem and its corollary.

Theorem 1.

$$P\left\{\sup_{\alpha\in\Lambda} |R(\alpha) - R_{emp}(\alpha)| > \epsilon\right\} \le 4\exp\left\{\left(\frac{H^\Lambda_{ann}(2\ell)}{\ell} - (\epsilon - \frac{1}{\ell})^2\right)\ell\right\} \tag{3}$$

for any distribution $F(z_1, \ldots z_\ell)$.

Corollary 1. *For the existence of nontrivial exponential bounds on uniform convergence,*

$$\lim_{\ell\to\infty} \frac{H^\Lambda_{ann}(\ell)}{\ell} = 0 \tag{4}$$

is sufficient.

Eventually we will define the *VC dimension* of a class of functions, which provides an upper bound on $H^\Lambda_{ann}(\ell)$. In fact, we have $H^\Lambda(\ell) \le H^\Lambda_{ann}(\ell) \le$ expressiondependingonVC Dimension. From $H^\Lambda(\ell)$ to $H^\Lambda_{ann}(\ell)$ to VC dimension, we have looser and looser bounds, but each one is more easily computable than the previous ones.

In order to prove Theorem 1, we will make use of the following basic lemma.

Basic Lemma. *For a given sample, $Z^{2\ell} \triangleq z_1, \ldots z_\ell, z_{\ell+1}, \ldots z_{2\ell}$, define the empirical risks on the first half and second half samples as*

$$\nu(\alpha, Z_1) \triangleq \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i, \alpha) \tag{5}$$

$$\nu(\alpha, Z_2) \triangleq \frac{1}{\ell}\sum_{i=\ell+1}^{2\ell} Q(z_i, \alpha) \tag{6}$$

The difference between the first and second half empirical risks is defined to be

$$\begin{aligned} \rho^\Lambda(\alpha, Z^{2\ell}) &\triangleq |\nu(\alpha, Z_1) - \nu(\alpha, Z_2)| && (7) \\ \rho^\Lambda(Z^{2\ell}) &\triangleq \sup_{\alpha\in\Lambda} \rho^\Lambda(\alpha, Z^{2\ell}) && (8) \end{aligned}$$

[1]All material adapted from Vapnik's Statistical Learning Theory, Chapter 4.

Lastly, define the different between empirical and true risk

$$\pi(\alpha, Z_1) \triangleq |R(\alpha) - \nu(\alpha, Z_1)| \tag{9}$$

$$\pi^\Lambda(Z_1) \triangleq \sup_{\alpha\in\Lambda} \pi^\Lambda(\alpha, Z_1) \tag{10}$$

The lemma states

$$P(\pi^\Lambda(Z_1) > \epsilon) \leq 2P(\rho^\Lambda(Z^{2\ell}) > \epsilon - \frac{1}{\ell}) \tag{11}$$

The lemma upper bounds fluctuations in π (maximum difference between empirical and real risks) by fluctuations in ρ (maximum difference between first and second half empirical risks). This upper bound is useful because ρ is much easier to compute than π. To calculate ρ, we can use the Law of Large Numbers and the triangle inequality; if two quantities are both close to the same mean, then they can't be very far apart.

Proof. By definition,

$$P(\rho^\Lambda(Z^{2\ell}) > \epsilon - \frac{1}{\ell}) = \int_{Z^{2\ell}} \theta(\rho^\Lambda(Z^{2\ell}) - \epsilon + \frac{1}{\ell}) dF(Z^{2\ell}) \tag{12}$$

Using Fubini's Theorem, which states that $\int_{x,y} dxdy = \int_x \int_y dydx$, we can separate the integral on the RHS into two integrals. We then lower bound the integral by fixing a "bad" set of Z_1's, and integrate w.r.t. Z_2 assuming Z_1 is fixed.

$$P(\rho^\Lambda(Z^{2\ell}) > \epsilon - \frac{1}{\ell}) = \int_{Z_1}\int_{Z_2} \theta(\rho^\Lambda(Z^{2\ell}) - \epsilon + \frac{1}{\ell}) dF(Z_2)dF(Z_1) \tag{13}$$

$$\geq \int_{Q}\int_{Z_2} \theta(\rho^\Lambda(Z^{2\ell}) - \epsilon + \frac{1}{\ell}) dF(Z_2)dF(Z_1) \tag{14}$$

$$\text{where} \quad Q \triangleq \{Z_1 : \pi^\Lambda(Z_1) > \epsilon\} \tag{15}$$

By definition, $\pi^\Lambda(Z_1) = \sup_{\alpha\in\Lambda} |R(\alpha) - \nu(\alpha, Z_1)|$. So $\pi^\Lambda(Z_1) > \epsilon$ implies that Z_1 gives the worst case fluctuation from the real risk. Hence there must exist α^* that achieves this ϵ fluctuation on Z_1.

$$|R(\alpha^*) - \nu(\alpha^*, Z_1)| > \epsilon \tag{16}$$

Let I denote the integrand of Equation (14). We have

$$I \triangleq \int \theta(\rho^\Lambda(Z^{2\ell}) - \epsilon + \frac{1}{\ell}) dF(Z_2) \tag{17}$$

$$= \int \theta \left[\sup_{\alpha\in\Lambda} \rho(\alpha, Z^{2\ell}) - \epsilon + \frac{1}{\ell}\right] dF(Z_2) \tag{18}$$

$$\geq \int \theta \left[\rho(\alpha^*, Z^{2\ell}) - \epsilon + \frac{1}{\ell}\right] dF(Z_2) \tag{19}$$

where in deriving the last inequality, we used the fact that $\rho(\alpha^*, Z^{2\ell}) < \sup_{\alpha\in\Lambda} \rho(\alpha, Z^{2\ell})$, therefore $\rho(\alpha^*, Z^{2\ell}) - \epsilon + \frac{1}{\ell} \geq 0$ is a more stringent condition than $\sup \rho(\alpha, Z^{2\ell}) - \epsilon + \frac{1}{\ell} \geq 0$. Integrating over the more stringent indicator function then gives us a lower bound.

Now let $\nu(\alpha^*, Z_1) < R(\alpha^*) - \epsilon$. This means that the error frequency of α^* on the set Z_1 is lower than the actual risk of α^*, therefore empirical risk on Z_1 is misleadingly good. The above condition implies that $|R(\alpha^*) - \nu(\alpha^*, Z_1)| > \epsilon$, therefore $Z_1 \in Q$. Now, in order for Z_2 to satisfy $|\nu(\alpha^*, Z_1) - \nu(\alpha^*, Z_2)| > \epsilon - \frac{1}{\ell}$, it suffices that $\nu(\alpha^*, Z_2) \geq R(\alpha^*) - \frac{1}{\ell}$. Therefore we can lower bound Equation (19) by integrating over this smaller set of Z_2's.

$$\begin{aligned}
I &\geq \int \theta(\nu(\alpha^*, Z_2) - R(\alpha^*) + \frac{1}{\ell}) dF(Z_2) && (20)\\
&= P(\nu(\alpha^*, Z_2) \geq R(\alpha^*) - \frac{1}{\ell}) && (21)\\
&= \sum_{\frac{k}{\ell} \geq R(\alpha^* - \frac{1}{\ell})} \left(\begin{array}{c} \ell \\ k \end{array} \right) R(\alpha^*)^k (1 - R(\alpha^*))^{\ell - k} && (22)\\
&\geq \frac{1}{2} && (23)
\end{aligned}$$

In going from Equation (21) to Equation (22) we have used the fact that $\nu(\alpha^*, Z_2)$ is just a sum of Bernoulli random variables with heads probability $R(\alpha^*)$. The last inequality is a simple consequence of the sum starting from below the mean. (The exact proof is left as an exercise for the reader.) This also explains why there has to be an extra $-\frac{1}{\ell}$ term in our results.

Finally, from Equation (14), we have

$$\begin{aligned}
P(\rho^\Lambda(Z^{2\ell} > \epsilon - \frac{1}{\ell}) &= \int_Q I dF(Z_1) && (24)\\
&\geq \frac{1}{2} \int_Q dF(Z_1) && (25)\\
&= \frac{1}{2} P(\pi^\Lambda(Z_1) > \epsilon) && (26)
\end{aligned}$$

which concludes our proof. □