Stats 300b: Theory of Statistics	**Winter 2018**

Lecture 9 – February 6

Lecturer: John Duchi *Scribes: Maxime Cauchois and Damian Pavlyshyn*

Warning: *these notes may contain factual errors*

Reading: van der Vaart 5.2, 19.1, 19.2

1 Uniform laws of large numbers

Definition 1.1. *Let $\mathcal{F}$ be a collection of functions $f\colon \mathcal{X} \to \mathbb{R}$. Then $\mathcal{F}$ satisfies a uniform law of large numbers (ULLN) if*

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{p} 0,$$

where $Pf = \int f dP$ and $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of the sample $\{X_1, \ldots, X_n\}$.

Example 1 (Glivenko-Cantelli theorem)**:** Let $\mathcal{F} = \{f(x) = \mathbf{1}\{x \le t\}, t \in \mathbb{R}\}$ so that $P_n f = P(X \le t)$ for some $t \in \mathbb{R}$. Then

$$\sup_{f \in \mathcal{F}} |P_n f - pf| = \sup_{t \in \mathbb{R}} |P_n(X \le t) - P(X \le t)| \xrightarrow{p} 0.$$

In fact, more is possible: the Dvoretzky-Kiefer-Wolfowitz inequality states that, for any $\epsilon > 0$,

$$\mathbb{P}\big(\sup_{t \in \mathbb{R}} |P_n(X \le t) - P(X \le t)| \ge \epsilon\big) \le 2\exp\{-2n\epsilon^2\}.$$

♣

Why do we want ULLNs? They make consistency results *much* easier. We'll give a few "generic" consistency results.

Let Θ be some parameter space, $\ell_\theta\colon \mathcal{X} \to \mathbb{R}$ some loss function, for example

$$\ell_\theta = -\log p_\theta(x)$$

for some model p_θ.

Then define the risk $R(\theta) = \mathbb{E}\ell_\theta(X) = P\ell_\theta$ and the observed risk $R_n(\theta) = P_n \ell_\theta$.

Observation 1 (Simple consistency results)**.** *If $\mathcal{F} = \{\ell_\theta\}_{\theta \in \Theta}$ satisfies a ULLN and $\{\hat{\theta}_n\}_n$ is any sequence of estimators such that*

$$R_n(\hat{\theta}_n) \le \inf_{\theta \in \Theta} R(\theta) + o_{\mathbb{P}}(1),$$

then $R(\hat{\theta}_n) \xrightarrow{p} \inf_\theta R(\theta)$.

Proof Assume w.l.o.g. that $\theta^* \in \operatorname{argmin}_\theta R(\theta)$. Then

$$
\begin{aligned}
R(\hat{\theta}_n) - R(\theta^*) &= \big(R(\hat{\theta}_n) - R_n(\hat{\theta}_n)\big) + \big(R_n(\hat{\theta}_n) - R_n(\theta^*)\big) + \big(R_n(\theta^*) - R(\theta^*)\big) \\
&= \underbrace{\sup_{\theta\in\Theta}|R_n(\theta) - R(\theta)|}_{o_{\mathbb{P}}(1) \text{ by ULLN}} + \underbrace{R_n(\hat{\theta}_n) - R_n(\theta^*)}_{o_{\mathbb{P}}(1) \text{ by assumption}} + \underbrace{R_n(\theta^*) - R(\theta^*)}_{o_{\mathbb{P}}(1) \text{ by regular LLN}} \\
&\xrightarrow{p} 0.
\end{aligned}
$$

□

Corollary 2 (Argmax/argmin theorem)**.** *Assume that R is such that, for all $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$R(\theta) \geq R(\theta^*) + \delta \textit{ whenever } d(\theta, \theta^*) \geq \epsilon.$$

Under the conditions of observation 1,

$$\hat{\theta}_n \xrightarrow{p} \theta^*.$$

Proof If $d(\hat{\theta}_n, \theta^*) \geq \epsilon$, then $R(\hat{\theta}_n) - R(\theta^*) \geq \delta$. But, by observation 1, $\hat{\theta}_n \xrightarrow{p} \theta^*$. Hence,

$$\mathbb{P}\big(d(\hat{\theta}_n, \theta^*) \geq \epsilon\big) \leq \mathbb{P}\big(R_n(\hat{\theta}_n) \geq R(\theta^*) + \delta\big) \to 0.$$

□

How do we prove ULLNs? Covering and understanding the "massiveness" of sets of functions.

Definition 1.2. *Let (Θ, ρ) be a (pseudo-)metric space.*

$$\rho\colon \Theta \times \Theta \to \mathbb{R}_{\geq 0}.$$

For $\epsilon > 0$, we say that $\{\theta^i\}_{i=1}^N$ is an ϵ-cover of Θ if, for all $\theta \in \Theta$, there exists an i such that

$$d(\theta, \theta^i) \leq \epsilon.$$

Definition 1.3. *The ϵ-covering number of Θ is the smallest size of ϵ-covers. ie,*

$$N(\Theta, \rho, \epsilon) = \inf\{N \in \mathbb{Z}_{\geq 0} : \textit{there exists an } \epsilon\textit{-cover } \{\theta^i\}_{i=1}^N \textit{ of } \Theta\}.$$

The metric entropy is then $\log N(\Theta, \rho, \epsilon)$.

Definition 1.4. *For $\delta > 0$, a set $\{\theta^i\}_{i=1}^N \subseteq \Theta$ is a δ-packing of Θ if, for all $i \neq j$*

$$\rho(\theta^i, \theta^j) > \delta.$$

The packing number is then

$$M(\Theta, \rho, \delta) = \sup\{M \in \mathbb{Z}_{\geq 0} : \textit{there exists a } \delta\textit{-cover } \{\theta^i\}_{i=1}^M \textit{ of } \Theta\}.$$

Observation 3. *For all $\epsilon > 0$,*

$$M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon).$$

Example 2 (Covering numbers of norm balls by volume arguments)**:** Let $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\| \leq r\}$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$ and $r > 0$.

Using $\rho(x, y) = \|x - y\|$, we have that

$$\left(\frac{r}{\epsilon}\right)^d \leq N(\Theta, \rho, \epsilon) \leq \left(1 + \frac{2r}{\epsilon}\right)^d.$$

Proof Observe that, for $\mathbf{B} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$, we have that

$$\frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(\epsilon\mathbf{B})} = \frac{\mathrm{Vol}(r\mathbf{B})}{\mathrm{Vol}(\epsilon\mathbf{B})} = \frac{r^d}{\epsilon^d}.$$

Hence, any covering of Θ must have at least $(r/\epsilon)^d$ ϵ-balls, and so

$$N(\Theta, \rho, \epsilon) \geq \left(\frac{r}{\epsilon}\right)^d.$$

To see the reverse inequality, let $\{\theta^i\}_{i=1}^M$ be a maximal ϵ-packing of $\Theta = r\mathbf{B}$. Then the $\theta^i + \mathbf{B}\epsilon/2$ are disjoint, and so

$$\biguplus_{i=1}^M \left(\theta^i + \frac{\epsilon}{2}\mathbf{B}\right) \subseteq \left(r + \frac{\epsilon}{2}\right)\mathbf{B}.$$

Therefore, we have that

$$\begin{aligned}
M(\epsilon/2)^d \,\mathrm{Vol}(\mathbf{B}) &= \sum_{i=1}^n \mathrm{Vol}(\theta^i + \mathbf{B}\epsilon/2) \\
&= \mathrm{Vol}\left(\biguplus_{i=1}^M (\theta^i + \mathbf{B}\epsilon/2)\right) \\
&\leq \mathrm{Vol}\big((r + \epsilon/2)\mathbf{B}\big) \\
&= (r + \epsilon/2)^d \,\mathrm{Vol}(\mathbf{B}).
\end{aligned}$$

Hence, we can conclude that

$$\begin{aligned}
M &\leq (2/\epsilon)^d (r + \epsilon/2)^d \\
&= (1 + 2r/\epsilon)^d.
\end{aligned}$$

♣

2 Bracketing number

When dealing with functional spaces $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$, a similar notion to covering numbers is the bracketing number, namely:

Definition 2.1. *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ be a collection of functions, and μ a measure on $\mathcal{X}$. A set $\{[l_i, u_i]\}_{i=1}^N \subset \mathbb{R}^{\mathcal{X}}$ is a ϵ - bracket of $\mathcal{F}$ in $L_p(\mu)$ if*

$$\forall f \in \mathcal{F}\ \exists i \text{ s.t } l_i \leq f(x) \leq \mu_i \text{ and } \|u_i - l_i\|_{L_p(\mu)} \leq \epsilon$$

From ϵ brackets, we similarly get bracketing numbers by taking the infimum over N:

Definition 2.2. *The bracketing number of $\mathcal{F}$ is*

$$N_{[]}\big(\mathcal{F}, L_p(\mu), \epsilon\big) := \inf \Big\{ N \in \mathbb{N} : \exists \text{ an } \epsilon\text{-bracket } \{[l_i, u_i]\}_{i=1}^N \text{ of } \mathcal{F} \text{ in } L_p(\mu) \Big\}$$

Example 3 (Lipschitz loss functions)**:** Let $\Theta \subset \mathbb{R}^d$ be compact, which implies that, for all $\epsilon > 0$, we have $N(\Theta, \|\cdot\|, \epsilon) < \infty$.

Let $\mathcal{F} = \{l_\theta : \theta \in \Theta\}$ where l_θ are $L(X)$-Lipschitz in θ, namely, for all x and θ_1, θ_2:

$$|l_{\theta_1}(x) - l_{\theta_2}(x)| \leq L(x) \|\theta_1 - \theta_2\|$$

Then, assuming that $\mathbb{E}[L(X)] < \infty$:

$$N_{[]}\left(\mathcal{F}, L_1, \epsilon\mathbb{E}[L(X)]\right) \leq N(\Theta, \|\cdot\|, \epsilon/2)$$

♣

Proof Let $\{\theta_i\}_{i=1}^N$ be an $\epsilon/2$ -covering of Θ, then let's define :

$$u_i(x) := l_{\theta_i}(x) + \frac{\epsilon}{2} L(x)$$
$$l_i(x) := l_{\theta_i}(x) - \frac{\epsilon}{2} L(x)$$

We know that for any $\theta \in \Theta$, $\exists \theta_i$ s,t $\|\theta - \theta_i\| \leq \frac{\epsilon}{2}$, and from Lipschitz properties of l_θ, we have:

$$\begin{aligned} |l_\theta(x) - l_{\theta_i}(x)| &\leq L(x) \|\theta - \theta_i\| \\ &\leq \frac{\epsilon}{2} L(x). \end{aligned}$$

Thus, for all $x \in \mathcal{X}$:

$$l_i(x) \leq l_\theta(x) \leq u_i(x)$$

As, for all $1 \leq i \leq N$, $\mathbb{E}[u_i(X) - l_i(X)] = \epsilon\mathbb{E}[L(X)]$, this ends the proof. □

3 Examples and theorems of uniform laws of large numbers

Theorem 4 (First ULLN)**.** *Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ satisfy:*

$$N_{[]}\left(\mathcal{F}, L_p, \epsilon\right) < \infty \text{ for all } \epsilon > 0$$

Then , under i.i.d. sampling

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{p} 0$$

Proof For any given $\epsilon > 0$, let $\{[l_i, u_i]\}_{i=1}^N$ be an ϵ-bracket for $\mathcal{F}$.

For any $f \in \mathcal{F}$, there exists $i \in [N]$ s.t $l_i \leq f \leq u_i$, and therefore we have:

$$\begin{aligned} P_n f - Pf &\leq P_n u_i - P l_i \\ &= P_n u_i - P u_i + P u_i - P l_i \\ &\leq (P_n - P) u_i + \epsilon. \end{aligned}$$

Similarly:

$$\begin{aligned} Pf - P_n f &\leq P u_i - P_n l_i \\ &\leq (P - P) l_i + \epsilon. \end{aligned}$$

This leads to:

$$\begin{aligned} \|P_n - P\|_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |P_n f - Pf| \\ &\leq \max_{1 \leq i \leq N} |(P_n - P)(u_i + l_i)| + \epsilon \\ &= o_p(1) + \epsilon \end{aligned}$$

as there are finitely many terms in the maximum. □

Example 4 (Logistic Regression)**:** Suppose that we are given pairs $Z = (X, Y) \in \mathbb{R}^d \times \{\pm 1\}$.

- <u>Goal</u>: Classification, find θ such that:

  $$\operatorname{sign}(\theta^T x) = y$$

- Consider the following loss function:

  $$l_\theta(x, y) = \log\left(1 + \exp(-y x^T \theta)\right)$$

  Then, considering that $\phi : t \mapsto \log(1 + \exp(-t))$ is 1-Lipschitz (its derivative being bounded by 1), we get that:

  $$|l_{\theta_1}(x, y) - l_{\theta_1}(x, y)| \leq \left|x^T(\theta_1 - \theta_2)\right| \leq \|x\| \, \|\theta_1 - \theta_2\|$$

  by Cauchy-Schwarz's inequality.

  Applying the result seen in the example 3 with $L(x) = \|x\|$, we see that, if $\Theta \subset \mathbb{R}^d$ is compact and X has a finite first moment, then:

  $$\|P_n - P\|_{\mathcal{F}} = \sup_{\theta \in \Theta} |P_n l_\theta - P l_\theta| \xrightarrow{p} 0$$

♣

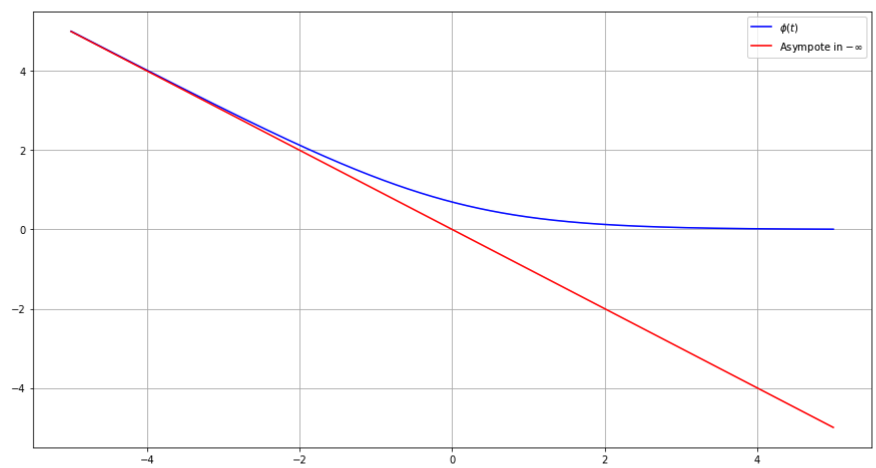

Figure 1: Loss function in logistic regression (in terms of $yx^T\theta$)