Stats 300b: Theory of Statistics **Winter 2019**

Lecture 9 – February 5

Lecturer: John Duchi *Scribe: Brett Larsen*

Warning: *these notes may contain factual errors*

Reading: van der Vaart 5.2, 19.1, 19.2

Outline: Uniform Laws of Large Numbers (ULLN)

- Argmax/argmin theorem
- Covering and bracketing numbers
- Metric entropies

1 Uniform laws of large numbers

Definition 1.1. *Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{F}$ satisfies a uniform law of large numbers (ULLN) for distribution P 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 $\{X_1, \ldots, X_n\}$.

For notational simplicity, we have defined the $\mathcal{F}$-norm of a measure as $\|\mu\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\int f \, d\mu|$. Then we have

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}[f(x)]|$$

Example 1 (Glivenko-Cantelli theorem)**:** Let $\mathcal{F} = \{f(x) = \mathbf{1}\{x \le t\}, t \in \mathbb{R}\}$ (the function class of step-down functions) 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 $\epsilon > 0$,

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

i.e., exponential concentration. ♣

Why do we want ULLNs? They make consistency results *much* easier. Let's consider a "generic" consistency results with a ULLN:

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

$$\ell_\theta = -\log p_\theta(x) \quad \text{where } p_\theta \text{ is the density of } X$$

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 $\{\widehat{\theta}_n\}_n$ is any sequence of estimators such that*

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

then $R(\widehat{\theta}_n) \xrightarrow{p} \inf_\theta R(\theta)$ (i.e the risk is consistent).

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

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

Note that we are ignoring issues of measurability. We will only work with separable metric spaces. □

Corollary 2 (Argmax/argmin theorem)**.** *Let R be such that for all $\epsilon > 0$ there exits some $\delta > 0$*

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

If $\mathcal{F} = \{\ell_\theta\}_{\theta\in\Theta}$ satisfies a ULLN observation and there exist a sequence of estimators that satisfy the conditions of observation 1, then

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

In other words, M-estimation yields consistent estimators.

Proof Pick any $\epsilon > 0$. If $d(\widehat{\theta}_n, \theta^*) \geq \epsilon$, then $R(\widehat{\theta}_n) \geq R(\theta^*) + \delta$. If $d(\widehat{\theta}_n, \theta^*) \geq 0$

$$\begin{aligned} \delta \leq R(\widehat{\theta}_n) - R(\theta^*) &\leq \sup_{\theta\in\Theta} |R_n(\theta) - R(\theta)| + \big(R_n(\widehat{\theta}_n) - R_n(\theta^*)\big) + \big(R_n(\theta^*) - R(\theta^*)\big) \\ &\leq o_{\mathbb{P}}(1) \end{aligned}$$

This is a contradiction $\implies d(\widehat{\theta}_n, \theta^*) \xrightarrow{p} 0$ as desired. □

Intuitivley, the above proof shows that the empricial risk loss surface being close to the risk loss surface means we can't move the minimizer too far away.

How do we prove ULLNs? Covering and understanding the "massiveness" of sets of functions. We will in general follow the process:

1. Choose certain exemplar functions $\{f_i\}$
2. Put balls around each function (functions that are similar in some metric)
3. If we can aruge that $|P_n f - Pf|$ is similar within each ball, then we only need to show convergence of the individual $\{f_i\}$

Definition 1.2. *Let (Θ, ρ) be a metric space (may also be a semi- or pseudo-metric):*

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

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

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

(Note for us it is note necessary that θ^i belong to Θ)

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

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

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

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

Observation 3. *(Relationship between packing and covering numbers) 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 $\rho(\theta, \theta') = \|\theta - \theta'\|$, we can bounde the covering number as follows:

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

Proof Let $\mathbf{B} = \{\theta \in \mathbb{R}^d : \|\theta\| \leq 1\}$ be the unit ball. Observe that we have:

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

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

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

Conversely, suppose $\{\theta^i\}_{i=1}^M$ is 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

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

Using the relationship between covering/pakcing numbers, we can obtain the other side of the inequality:

$$N(\epsilon) \leq M(\epsilon) \leq \frac{(r+\epsilon/2)^d}{\left(\frac{\epsilon}{2}\right)^d} = \left(1 + \frac{2r}{\epsilon}\right)^d$$

Finally, in terms of metric entropy, we have obtained the following bound:

$$d \log \frac{r}{\epsilon} \leq \log N(\epsilon) \leq d \log \left(1 + \frac{2r}{\epsilon}\right)$$

♣

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 $\mathcal{F} \subseteq \left\{f : \mathcal{X} \to \mathbb{R}\right\}$ be a collection of functions, and μ a measure on $\mathcal{X}$. A set*

$$\{[l_i, u_i]\}_{i=1}^N \subset \{\mathcal{X} \to \mathbb{R}\}$$

(i.e. $\ell_i, u_i : \mathcal{X} \to \mathbb{R}$) is a ϵ - bracket of $\mathcal{F}$ in $L_p(\mu)$, $p \geq 1$ if

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

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

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

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

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

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

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

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

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

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

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 m_θ, we have:

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

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

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

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

Remark In the previous example, we have used Lipshitz functions combined with a compact parameter sapce to get bracketing numbers in L_1. Generally, if $\mathbb{E}[L_p(X)] < \infty$, we can get control over $N_{[]}(\mathcal{F}, L_p(\mu), \epsilon)$.

3 Examples and theorems of uniform laws of large numbers

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

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

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

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

Similarly:

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

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. □