Uniform Laws of Large Numbers

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- Uniform laws of large numbers
- "argmax" theorem
- Covering and bracketing numbers
- Metric entropy

Reading:

- van der Vaart Chapters 5.2, 19.1, 19.2.
- Wainwright Chapters 4, 5.1 cover the material but rely on some concentration inequalities we will cover in coming lectures.

Uniform laws of large numbers

Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{F}$ satisfies a *ULLN* (for a distribution P) if

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |P_n f - Pf| \stackrel{p}{\to} 0.$$

Example (Glivenko Cantelli)

Let $\mathcal{F} = \{f(x) = 1\{x \le t\}\}_{t \in \mathbb{R}}$. Then

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

More is possible: Dvoretzky-Kiefer-Wolfowitz inequality gives

$$\mathbb{P}\left(\sup_t |P_n(X \le t) - P(X \le t)| \ge \epsilon\right) \le 2\exp\left(-2n\epsilon^2\right).$$

Consistency and argmax theorems

- ULLNs make consistency results much easier
- easy "generic" consistency result for loss minimization
- Θ is a parameter space, $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$ a loss
- population loss (risk) $L(\theta) = P\ell(\theta, X)$ and $L_n(\theta) = P_n\ell(\theta, X)$

Proposition

If $\mathcal{F} = \{\ell(\theta, \cdot)\}_{\theta \in \Theta}$ satisfies the ULLN and

$$L_n(\widehat{\theta}_n) \le \inf_{\theta \in \Theta} L_n(\theta) + o_P(1) \quad \textit{then} \quad L(\widehat{\theta}_n) \xrightarrow{p} \inf_{\theta \in \Theta} L(\theta)$$

The argmax theorem

- Assume for all $\epsilon > 0$, there is $\delta > 0$ such that

$$L(\theta) \geq L(\theta^\star) + \delta \quad \text{whenever } d(\theta, \theta^\star) \geq \epsilon$$

Proposition (Argmax)

If $L_n(\widehat{\theta}_n) \leq \inf_{\theta \in \Theta} L_n(\theta) + o_P(1)$ *and* $\{\ell(\theta, \cdot)\}_{\theta \in \Theta}$ *satisfies the ULLN, then*

$$\widehat{\theta}_n \xrightarrow{p} \theta^\star.$$

Covering a set

Definition

Let (Θ, ρ) be a metric space (ρ may be a semimetric). For $\epsilon > 0$, a set $\{\theta^i\}_{i=1}^N$ is an ϵ-*cover* of Θ if for each $\theta \in \Theta$ there exists $i \leq N$ such that

$$\rho(\theta, \theta^i) \leq \epsilon$$

- Sometimes require $\theta^i \in \Theta$, in which case we have *internal cover*

Packing a set

Definition

For $\delta > 0$, a set $\{\theta^i\}_{i=1}^M \subset \Theta$ is a δ-*packing* of Θ is $\rho(\theta^i, \theta^j) > \delta$ for each $i \neq j$.

Covering numbers and entropies

Definition

The ϵ-*covering* number $N(\Theta, \rho, \epsilon)$ of Θ is the smallest N such that there exists an ϵ-cover $\{\theta^i\}_{i=1}^N$ of Θ.

Definition

The δ-*packing number* $M(\Theta, \rho, \delta)$ of Θ is the largest M such that there exists a δ-packing $\{\theta^i\}_{i=1}^M \subset \Theta$ of Θ.

Definition (Entropies)

The metric entropy of Θ is $\log N(\Theta, \rho, \epsilon)$; the packing entropy of Θ is $\log M(\Theta, \rho, \epsilon)$.

Proposition (Equivalence between entropies)

$$M(\Theta, \rho, 2\epsilon) \leq N(\Theta, \rho, \epsilon) \leq M(\Theta, \rho, \epsilon).$$

Covering numbers by volume arguments

Let $\mathbb{B}^d = \{\theta \in \mathbb{R}^d \mid \|\theta\| \le 1\}$ be the 1-ball for norm $\|\cdot\|$.

Proposition (Entropy of norm balls)

For any $0 < \epsilon \le r < \infty$,

$$d \log \frac{r}{\epsilon} \le \log N(r\mathbb{B}^d, \|\cdot\|, \epsilon) \le d \log \left(1 + \frac{2r}{\epsilon}\right).$$

Bracketing numbers

- for $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$, an additional type of covering is useful

Definition

Let $\mathcal{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$, μ a measure on $\mathcal{X}$, and $p \geq 1$. A set $\{[l_i, u_i]\}_{i=1}^N \subset L^p(\mu)$ is an *ϵ-bracketing of $\mathcal{F}$ for $L^p(\mu)$* if for each $f \in \mathcal{F}$, there exists $i \in [N]$ satisfying

$$l_i \leq f \leq u_i \quad \text{and} \quad \|l_i - u_i\|_{L^p(\mu)} := \left(\int |l_i - u_i|^p d\mu\right)^{1/p} \leq \epsilon.$$

The *bracketing number* $N_{[]}(\mathcal{F}, L^p(\mu), \epsilon)$ of $\mathcal{F}$ is the smallest N such that there exists such an ϵ-bracket of size N.

Bracketing a parametric collection functions

- $\Theta \subset \mathbb{R}^d$ is compact with $N(\Theta, \|\cdot\|, \epsilon) < \infty$
- criterion functions $\ell_\theta(x)$ are $M(x)$-Lipschitz in θ with $\mathbb{E}[M(X)] < \infty$, i.e. $|\ell_{\theta_0}(x) - \ell_{\theta_1}(x)| \le M(x) \|\theta_0 - \theta_1\|$
- function class $\mathcal{F} = \{\ell_\theta\}_{\theta \in \Theta}$

Proposition

The bracketing number of $\mathcal{F}$ satisfies

$$N_{[]}(\mathcal{F}, L^1(P), \epsilon PM(X)) \le N(\Theta, \|\cdot\|, \epsilon/2).$$

A uniform law of large numbers

Theorem

Let $\mathcal{F} \subset \{\mathcal{X} \to \mathbb{R}\}$ *satisfy* $N_{[]}(\mathcal{F}, L^1(P), \epsilon) < \infty$ *for all* $\epsilon > 0$. *Then*

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

Example: logistic regression

- data $\{x, y\} \in \mathbb{R}^d \times \{\pm 1\}$
- losses $\ell(\theta, x, y) = \log(1 + \exp(-y\langle\theta, x\rangle))$
- function class $\mathcal{F}_{\log} = \{\ell(\theta, \cdot)\}_{\theta \in \Theta}$

Proposition

If $P\|X\| < \infty$, then $\|P_n - P\|_{\mathcal{F}_{\log}} \overset{p}{\to} 0$.