Uniform Law of Large Numbers

Marcelo Ortiz-Villavicencio

February 2, 2024

Econometrics Reading Group

Roadmap

1. Introduction
   - Uniform convergence of CDF
   - Uniform law for more general function classes
2. A uniform law via Rademacher Complexity
3. Upper bounds on the Rademacher complexity
   - VC dimension

Introduction

Introduction

- In the previous chapter we pointed out some limitations of asymptotic analysis in high dimensions.
- In this chapter, we will turn our focus to *non-asymptotic analysis*, where we provide convergence guarantees without having the number of observations $n \to \infty$.
- Our focus in this lecture is a set of results called the *uniform law of large numbers*.
- These results represent a strengthening of the usual LLN, which applies to a fixed sequence of RV, to related laws that hold *uniformly* over collections of RV.
- Such uniform laws are of theoretical interest in *Empirical Process Theory*.
- Moreover, they also play an important role in understanding the behavior of different statistical estimators providing guarantees for bounds of the form:

$$\Pr\left[\sup_{h \in \mathcal{H}} |\hat{\mathcal{L}}(h) - \mathcal{L}(h)| \leq \epsilon\right] \geq 1 - \delta.$$

Some terminology from Empirical Process Theory

EMORY UNIVERSITY

- A *stochastic process* is a collection of random variables $\{X(t), t \in T\}$ on the same probability space, indexed by an arbitrary set T.
- An *empirical process* is a stochastic process based on a random sample.
- Consider a random sample $X_1, \ldots, X_n$ of independent draws from a probability measure P on an arbitrary sample $\mathcal{X}$.
- For a set A, we define the *empirical measure* (distribution) of A to be

$$\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A)$$

  where $\delta_x(A)$ is the *Dirac measure* (or point mass) that assigns mass 1 if $x \in A$ and zero elsewhere
- An *Indicator function* is closely related to a Dirac.

Some terminology from Empirical Process Theory

- Given some integrable function g, we may define the expectation functional γ_g via
$$\gamma_g(P) = \int g(x)dP(x) = \mathbb{E}[g(X)]$$
- We can think about the previous expression as its an *empirical integral*
$$\gamma_g(\mathbb{P}_n) = \int g(x)d\mathbb{P}_n(x) = \mathbb{E}_n[g(X)]$$
- For any class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$, an *empirical process* $\{\gamma_f(\mathbb{P}_n), f \in \mathcal{F}\}$ can be defined.
- Our goal is to show "how close" $\{\gamma_f(\mathbb{P}_n), f \in \mathcal{F}\}$ is to $\{\gamma_f(P), f \in \mathcal{F}\}$.

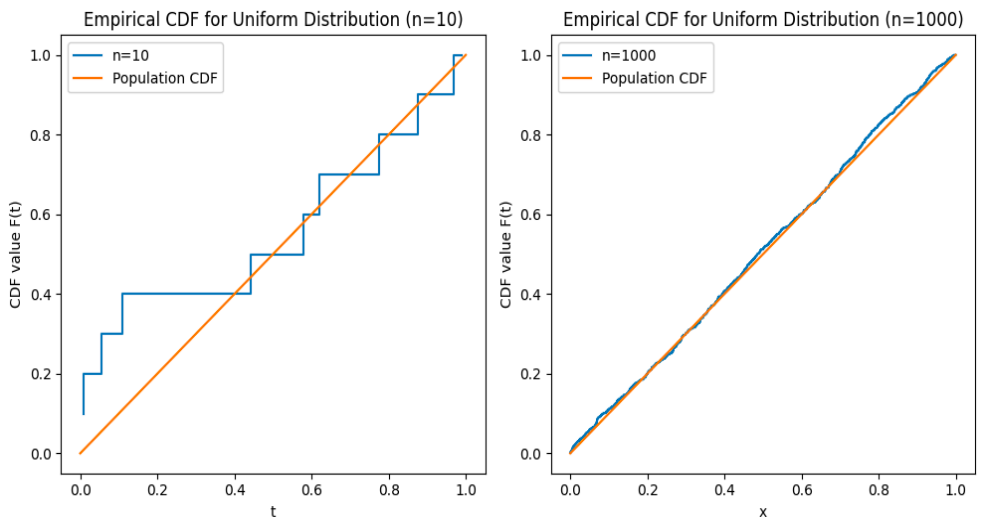
Uniform convergence of CDF

- The law of any scalar random variable X can be fully specified by its *cumulative distribution function* (CDF), whose value at any point $t \in \mathbb{R}$ is given by $F(t) := P(X \leq t)$.
- Suppose a collection $\{X_i\}_{i=1}^n$ of n i.i.d. samples, each drawn according to the law specified by F.
- A natural estimate of F is the empirical CDF given by

$$\widehat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, n]}[X_i]$$

  where $1_{(-\infty,t)}[x]$ is a $\{0,1\}$-valued *indicator function* for the event $\{x \leq t\}$.
- Since the population CDF can be written as $F(t) = \mathbb{E}\left[1_{(-\infty,t)}[X]\right]$, the empirical CDF is an *unbiased estimate*.

Uniform convergence of CDF



Glivenko-Cantelli

- Given a pair of CDFs F and G, let us measure the distance between them using the *sup-norm*

$$||G - F||_\infty := \sup_{t \in \mathbb{R}} |G(t) - F(t)|$$

- We can define then the *continuity* of a functional γ with respect to this norm.
- More precisely, the functional γ is *continuous* at F in the sup-norm if, $\forall \epsilon > 0$, there exists a $\delta > 0$ such that $||G - F||_\infty \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \epsilon$

Theorem (Glivenko-Cantelli)

For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that

$$||\widehat{F}_n - F||_\infty \xrightarrow{a.s.} 0.$$

- This notion is useful because, for *any* continuous functional, it reduces the consistency question to the issue of whether or not that difference converges to zero.

Uniform law for more general function classes

- Let's focus on a more general consideration of ULLN.
- Let $\mathscr{F}$ be a class of integrable real-valued functions with domain X, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$.
- Consider the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right|,$$

  which measures the absolute deviation between the sample average $\frac{1}{n}\sum_{i=1}^n f(X_i)$ and the population average $\mathbb{E}[f(X)]$, uniformly over the class $\mathscr{F}$.

Theorem

We say that $\mathscr{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ converges to zero in probability as $n \to \infty$.

Note: Not all classes of functions are Glivenko-Cantelli.

In what other statistical contexts do these quantities arise?

EMORY UNIVERSITY

- These quantities are one of the main focus of methods based on *empirical risk minimization*.
- Consider an indexed family of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$,
- Suppose we have access to n samples $\{X_i\}_{i=1}^n$, each sample lying in some space $\mathcal{X}$.
- Those samples are drawn i.i.d. according to a distribution $\mathbb{P}_{\theta^*}$, for some *fixed* but *unknown* $\theta^* \in \Omega$. Here the index θ^* could lie within a *finite-dimensional space*, such as $\Omega = \mathbb{R}^d$, or could lie within some function class $\Omega = \mathscr{G}$, in which case the problem is nonparametric (i.e., *infinite-dimensional space*).
- A standard *decision-theoretic* approach to estimating θ^* is based on minimizing a **cost function** of the form $\theta \mapsto \mathcal{L}_\theta(X)$, which measures the "fit" between a parameter $\theta \in \Omega$ and the sample $X \in \mathcal{X}$.

Empirical Risk Minimization

- Given the collection of n samples $\{X_i\}_{i=1}^n$, the principle of *empirical risk minimization* is based on the objective function (a.k.a. *empirical risk*)

$$\widehat{R}_n(\theta, \theta^*) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\theta(X_i)$$

- The empirical risk is contrasted with the *population risk*, where the expectation $\mathbb{E}_{\theta^*}$ is taken over a sample $X \sim \mathbb{P}_{\theta^*}$,

$$R(\theta, \theta^*) := \mathbb{E}_{\theta^*}[\mathcal{L}_\theta(X)],$$

- In practice, one minimizes the empirical risk over some *subset* Ω_0 of the full space Ω, thereby obtaining some estimate $\widehat{\theta}$.
- The question is how to bound the *excess risk*, measured in terms of the population quantities - namely the difference

$$E(\widehat{\theta}, \theta^*) := R(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega_0} R(\theta, \theta^*).$$

How to bound the excess risk?

- Our goal is to develop methods for controlling the *excess risk*.
- Suppose there exists some $\theta_0 \in \Omega_0$ such that

$$R(\theta_0, \theta^*) = \inf_{\theta \in \Omega_0} R(\theta, \theta^*)$$

- Then, the excess risk can be decomposed as

$$E(\widehat{\theta}, \theta^*) = \underbrace{\{R(\widehat{\theta}, \theta^*) - \widehat{R}_n(\widehat{\theta}, \theta^*)\}}_{T_1} + \underbrace{\{\widehat{R}_n(\widehat{\theta}_0, \theta^*) - \widehat{R}_n(\theta_0, \theta^*)\}}_{T_2 \leq 0} + \underbrace{\{\widehat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)\}}_{T_3}$$

- T_2 is non-positive, since $\hat{\theta}$ minimizes the empirical risk over Ω_0.
- Because θ_0 is an unknown but non-random quantity, and recalling the definition of empirical risk, T_3 can be rewritten as

$$T_3 = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\theta_0}(X_i)\right] - \mathbb{E}_X\left[\mathcal{L}_{\theta_0}(X)\right]$$

corresponding to the deviation of a sample mean from its expectation for the random variable $\mathcal{L}_{\theta_0}(X)$.

How to bound the excess risk?

- T_1 can be written in a similar way, namely as the difference

$$T_1 = \mathbb{E}_X\left[\mathcal{L}_{\hat{\theta}}(X)\right] - \left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\hat{\theta}}\left(X_i\right)\right].$$

- This quantity is more challenging to control, because the parameter $\widehat{\theta}$ (in contrast to θ_0) is now random, and depends on the samples $\{X_i\}_{i=1}^n$.
- Hence, controlling T_1 requires a *stronger* result, such as a *uniform law of large numbers* over the cost function class $\mathscr{L}\left(\Omega_0\right) := \{x \mapsto \mathcal{L}_\theta(x), \theta \in \Omega_0\}$.
- With this notation, we have

$$T_1 \leq \sup_{\theta\in\Omega_0}\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_\theta\left(X_i\right) - \mathbb{E}_X\left[\mathcal{L}_\theta(X)\right]\right| = \left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathcal{L}(\Omega_0)}$$

- This same quantity also dominates T_3, we conclude that the *excess risk* is at most

$$2\left\|\mathbb{P}_n - \mathbb{P}\right\|_{\mathfrak{L}(\Omega_0)}$$

A uniform law via Rademacher Complexity

Uniform Law via Rademacher Complexity

- Let me now turn to the technical details of deriving such results.
- An important concept that underlies the study of uniform laws is the *Rademacher complexity* of the function class $\mathscr{F}$.
- For any fixed collection $x_1^n := (x_1, x_2, \cdots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathscr{F}(x_1^n) := \{(f(x_1), f(x_2), \cdots, f(x_n)) \mid f \in \mathscr{F}\}$$

- The set $\mathscr{F}(x_1^n)$ corresponds to all those vectors in $\mathbb{R}^n$ that can be realized by applying a function $f \in \mathscr{F}$ to the collection $(x_1, x_2, \cdots, x_n)$and the *empirical Rademacher complexity* is given by

$$\mathcal{R}\left(\mathscr{F}\left(x_1^n\right)/n\right) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right| \right].$$

- Given a collection $X_1^n := \left\{X_i\right\}_{i=1}^n$ of random samples, then the *empirical Rademacher complexity* $\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)$ is a random variable.

Rademacher Complexity

- Taking its expectation yields the Rademacher complexity of the function class $\mathscr{F}$ we get

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_X\left[\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)\right] = \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right].$$

- Note that the *Rademacher complexity* is the average of the maximum correlation between the vector $(f(X_1),\ldots,f(X_n))$ and the "noise vector" $(\varepsilon_1,\ldots,\varepsilon_n)$, where the maximum is taken over all functions $f \in \mathscr{F}$.
- Intuition: a function class is extremely large if we can always find a function that has a high correlation with a randomly drawn noise vector. Conversely, when the Rademacher complexity decays as a function of sample size, then it is impossible to find a function that correlates very highly in expectation with a randomly drawn noise vector.
- Simple words: If the $\mathcal{R}_n(\mathscr{F})$ is small, it suggests that the function class is not very sensitive to random noise in the data. In other words, small $\mathcal{R}_n(\mathscr{F})$ often implies better generalization performance in a learning algorithm.

Rademacher complexity & Glivenko-Cantellii

- There is a connection between *Rademacher complexity* and the *Glivenko-Cantelli* theorem.
- In particular, any bounded function class $\mathscr{F}$, the condition $\mathcal{R}_n(\mathscr{F}) = o(1)$ implies the *Glivenko-Cantelli* property.

Theorem

For any b-uniformly bounded class of functions $\mathscr{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \leq 2\mathcal{R}_n(\mathscr{F}) + \delta$$

with probability at least $1 - \exp\left(-\frac{n\delta^2}{2b^2}\right)$. Consequently, as long as $\mathcal{R}_n(\mathscr{F}) = o(1)$, we have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{a.s.} 0$.

- This is nothing more than a tail bound for the probability that the RV $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ deviates *substantially above* a multiple of the Rademacher complexity.
- Therefore, we need to obtain upper bounds on the $\mathcal{R}_n(\mathscr{F})$

Upper bounds on the Rademacher complexity

Upper bounds on the Rademacher complexity

- To make the previous theorem useful, we require methods for *upper bounding* the Rademacher complexity.
- There are several methods to do so, ranging from simple union bounds (suitable for finite function classes) to more advanced techniques involving *metric entropy* and *chaining* (I will skip this due to time constraints, sorry).
- Instead, we gonna focus on more ”elementary” techniques that apply for function classes with *polynomial discrimination* and *Vapnik-Chervonenski* classes.

Polynomial discrimination

- It is relatively straightforward to establish uniform laws for function classes with *polynomial discrimination*
- Our interest in function classes for which the *cardinality* grows only as a polynomial function of sample size.

Definition (Polynomial discrimination)

A class $\mathscr{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $v \geq 1$ if, for each positive integer n and collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathcal{X}$, the set $\mathscr{F}(x_1^n)$ has *cardinality upper bounded*

$$\text{card}\left(\mathscr{F}(x_1^n)\right) \leq (n+1)^v$$

Controlling Rademacher Complexity

- Previous property provides a straightforward approach to controlling the Rademacher complexity.
- For any set $S \subset \mathbb{R}^n$, we use $D(S) := \sup_{x \in S} \|x\|_2$ to denote its *maximal width* in the ℓ_2-norm.

Lemma (Upper bound on Rademacher Complexity)

Suppose that $\mathscr{F}$ has polynomial discrimination of order v. Then for all positive integers n and any collection of points $x_1^n = (x_1, \ldots, x_n)$,

$$\underbrace{\mathbb{E}_\varepsilon \left[\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]}_{\mathcal{R}\left(\mathscr{F}\left(x_1^n\right)/n\right)} \leq 4D\left(x_1^n\right) \sqrt{\frac{v \log(n+1)}{n}},$$

where $D\left(x_1^n\right) := \sup_{f \in \mathscr{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2-radius of the set $\mathscr{F}\left(x_1^n\right)/\sqrt{n}$.

Simple case: Bounded functions

- A special simple case is when the function class is b *uniformly bounded* so that $D(x_1^n) \leq b$ for all samples.
- Applying the lemma

$$\mathcal{R}_n(\mathscr{F}) \leq 2b\sqrt{\frac{v\log(n+1)}{n}} \text{ for all } n \geq 1$$

- Combined with the Theorem, we conclude that any bounded function class with polynomial discrimination is Glivenko-Cantelli.
- What types of function classes have polynomial discrimination? A good example is based on *indicator functions* of the *left-sided intervals* $(-\infty, t]$ (e.g., CDFs)

Corollary (Classical Glivenko-Cantelli)

Let $F(t) = \mathbb{P}[X \leq t]$ be the CDF of a random variable X, and let $\widehat{F}_n$ be the empirical CDF based on n i.i.d. samples $X_i \sim \mathbb{P}$. Then

$$\mathbb{P}\left[\left\|\widehat{F}_n - F\right\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \leq e^{-\frac{n\delta^2}{2}} \quad \text{for all } \delta \geq 0,$$

and hence $\left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0.$

Vapnik-Chernonenkis dimension

- In this section, we briefly discuss a classical notion of complexity measure of function class, VC dimension.
- Let us consider a function class $\mathscr{F}$ in which each function f is binary-valued, taking the values $\{0, 1\}$ for concreteness.
- In this case, the set $\mathscr{F}(x_1^n)$ can have at most 2^n elements.

Definition (Shattering and VC dimension)

Given a class $\mathscr{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \ldots, x_n)$ is shattered by $\mathscr{F}$ if $\operatorname{card}(\mathscr{F}(x_1^n)) = 2^n$. The VC dimension $v(\mathscr{F})$ is the largest integer n for which there is some collection $x_1^n = (x_1, \ldots, x_n)$ of n points that is shattered by $\mathscr{F}$.

- When the quantity $v(\mathscr{F})$ is finite, then the function class $\mathscr{F}$ is said to be a *VC* class.
- Let's finish with an example.

VC dimension in supervised learning

- This example was taken from Ma (2022)
- Will show that VC dimension is an upper bound on the Rademacher complexity.
- The labels belong to the output space $\mathcal{Y} = \{-1, 1\}$, each classifier is a function $h : \mathcal{X} \to \mathbb{R}$ for all $h \in \mathcal{H}$, and the prediction is the sign of the output, i.e. $\hat{y} = \text{sgn}(h(x))$.
- We will look at zero-one loss function, i.e. $\ell_{0-1}((x, y), h) = 1(\text{sgn}(h(x)) \neq y)$. Note that we can re-express the loss function as

$$\ell_{0-1}((x, y), h) = \frac{1 - \text{sgn}(h(x))y}{2}.$$

VC dimension in supervised learning

- Think about the Rademacher complexity of ℓ_{0-1} loss function, i.e. considering the family of functions $\mathscr{F} = \{z = (x, y) \mapsto \ell_{0-1}((x, y), h) : h \in \mathcal{H}\}$.
- Define Q to be the set of all possible outputs on our dataset: $Q = \left\{\left(\text{sgn}\left(h\left(x^{(1)}\right)\right), \ldots, \text{sgn}\left(h\left(x^{(n)}\right)\right)\right) \mid h \in \mathcal{H}\right\}$.
- Computing the *Rademacher Complexity* we have

$$\begin{aligned}\mathcal{R}_n(\mathscr{F}) &= \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \frac{1 - \text{sgn}(h(x^{(i)}))y_i}{2}\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sup_{v\in Q} \frac{1}{n}\langle \varepsilon, v\rangle\right]\end{aligned}$$

- For any particular $v \in Q$, notice that $\langle \varepsilon, v\rangle$ is a sum of bounded random variables, so we can use *Hoeffding's inequality* to obtain

$$\Pr\left[\frac{1}{n}\langle \varepsilon, v\rangle \geq t\right] \leq \exp\left(-nt^2/2\right)$$

VC dimension in supervised learning

- Taking the union bound over $v \in Q$, we see that

$$\Pr\left[\exists v \in Q \text{ such that } \frac{1}{n}\langle \varepsilon, v\rangle \geq t\right] \leq |Q| \exp\left(-nt^2/2\right).$$

- Thus, with probability at least $1-\delta$, it is true that

$$\sup_{v \in Q} \frac{1}{n}\langle v, \varepsilon\rangle \leq \sqrt{\frac{2(\log|Q| + \log(2/\delta))}{n}}$$

- Similarly, we can show that $\mathbb{E}\left[\sup_{v\in Q} \frac{1}{n}\langle v, \varepsilon\rangle\right] \leq O\left(\sqrt{\frac{\log|Q|+\log(2/\delta)}{n}}\right)$ holds.
- *VC dimension* is one way to deal with bounding the size of Q.
- However, it has some limitations because will always end up with a bound that depends somehow on the dimension.

Thanks!

marcelo.ortiz@emory.edu
marcelortiz.com
@marcelortizv

References

- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.
- Ma, T. (2022). *Lecture Notes for Machine Learning Theory (CS229M/STATS214)*.