Statistics 210B Lecture 8 Notes

Daniel Raban

February 10, 2022

1 Introduction to Empirical Process Theory

1.1 Convergence of CDFs and the Glivenko-Cantelli theorem

Let $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} X$. X has CDF $F(t)$, i.e.

$$F(t) = \mathbb{P}(X \leq t).$$

We can also define the **empirical CDF**

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}.$$

This is the CDF of the empirical distribution of the X_i.

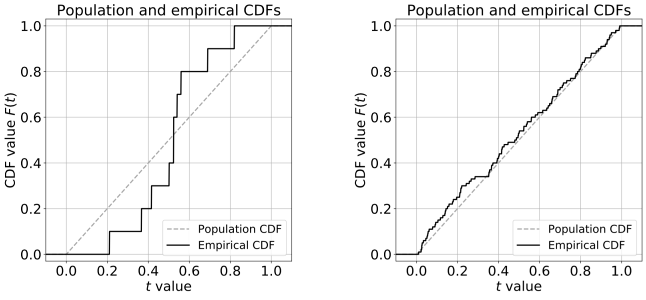
For any fixed t, the strong law of large numbers tells us that

$$\lim_{n\to\infty} \widehat{F}_n(t) = F(t) \qquad a.s.$$

If we are more ambitious, we may want convergence of functions. In this case, we look at the maximum difference,

$$\|F_n - F\|_\infty := \sup_{t\in[0,1]} |\widehat{F}_n(t) - F(t)|.$$

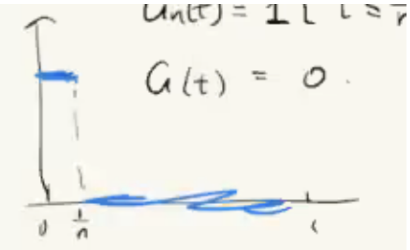
Here is a picture from Wainwright's book illustrating convergence of the empirical CDF to the uniform distribution on $[0,1]$.



Why is convergence of the supremum norm stronger than pointwise convergence? In general,

$$\lim_{n\to\infty} G_n(t) = G(t)\ \forall t \not\Longrightarrow \lim_{n\to\infty} \sup_t |G_n(t) - G(t)| = 0.$$

Example 1.1. Take $G_n(t) = \mathbb{1}_{\{t \le 1/n\}}$.



Then for any $t > 0$,$G_n(t) \to 0$, but $\lim_{n\to\infty} \sup_t |G_n(t) - G(t)| = \infty$.

A classical result guarantees uniform convergence of the empirical CDF.

Theorem 1.1 (Glivenko-Cantelli, 1933)**.** *Let* $X_i \overset{\text{iid}}{\sim} X$*, where* $F(t)$ *is the CDF of* X*. Then*

$$\lim_{n\to\infty} \|\widehat{F}_n - F\|_\infty = 0 \qquad a.s.$$

We will not prove this result. Instead, we will use empirical process theory, combined with concentration results to show something stronger:

$$\mathbb{P}\left(\|\widehat{F}_n - F\|_\infty \ge 8\sqrt{\frac{\log(n+1)}{n}} + t \right) \le \exp\left(-\frac{nt^2}{2}\right).$$

In other words,

$$\|\widehat{F}_n - F\|_\infty \le 8\sqrt{\frac{\log(n+1)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{with probability } 1-\delta.$$

Why is this result stronger? If we let $n \to \infty$, we get convergence in probability. We can get a.s. convergence using the Borel-Cantelli lemma.

1.2 Uniform laws for more general function classes

Suppose $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} X \sim \mathbb{P}$, and suppose we have a **function class** $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R} : \mathbb{E}[|f(X)|] < \infty\}$.

Definition 1.1. The **empirical process** indexed by $\mathcal{F}$ is

$$\left\{ \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right) : f \in \mathcal{F} \right\}.$$

Define

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

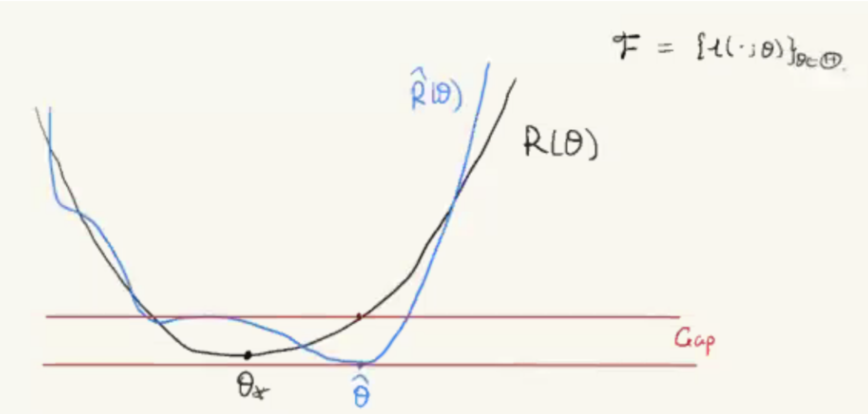
Here, $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ is the **empirical measure**. This is the object we will study for the next portion of the course. If there is only 1 function f, we can deal with this using the law of large numbers and concentration inequalities. We will learn how to deal with this object using empirical process theory.

Why do we care about the maximum of empirical process in statistics and machine learning? Recall the following setup:

Data distribution	$(X_i)_{i \in [n]} \overset{\text{iid}}{\sim} \mathbb{P}$
Loss function	$L : \mathcal{X} \times \Theta \to \mathbb{R}$
Empirical risk	$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \theta)$
Population risk	$R(\theta) = \mathbb{E}_{X \sim \mathbb{P}}[\ell(X; \theta)]$
Empirical risk minimizer	$\widehat{\theta} = \arg\min_{\theta} \widehat{R}(\theta)$
Population risk minimizer	$\theta_* = \arg\min_{\theta} R(\theta)$
Excess risk	$E = R(\widehat{\theta}) - R(\theta_*)$

We train $\widehat{\theta}$ on the empirical risk, so we want the empirical risk to be close to the population risk. So to make sure training on our training data is accurate, we want to make the excess risk small. The excess risk has the following decomposition:

$$E = \underbrace{(R(\widehat{\theta}) - \widehat{R}_n(\widehat{\theta}))}_{\text{Gap}} + \underbrace{(\widehat{R}_n(\widehat{\theta}) - \widehat{R}_n(\theta_*))}_{\leq 0} + \underbrace{(\widehat{R}_n(\theta_*) - R(\theta_*))}_{\text{bound using Hoeffding}}$$



The Gap is

$$\text{Gap} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(X; \widehat{\theta}) - \ell(X_i; \widehat{\theta})].$$

We cannot use the strong law of large numbers to examine this because the $\ell(X_i; \widehat{\theta})$ are not independent random variables. We can fix this by replacing $\widehat{\theta}$ by the sup over θ:

$$\leq \sup_{\theta} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(X;\theta) - \ell(X_i;\theta)] \right|.$$

Here, $f(X) = \ell(X;\theta)$, so we want to look at the function class $\mathcal{F} = \{\ell(\cdot;\theta) : \theta \in \Theta\}$.

Definition 1.2. We say that $\mathcal{F}$ is a **Glivenko-Cantelli class** for $\mathbb{P}$ if

$$\|P_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{p} 0.$$

Example 1.2. The Glivenko-Cantelli theorem says that $\mathcal{F}_1 = \{\mathbb{1}_{\{x \leq t\}}\}_{t \in \mathbb{R}}$ is a Glivenko-Cantelli class for any $\mathbb{P} \in \mathcal{P}(\mathbb{R})$.

Example 1.3. Consider $\mathcal{F}_2 = \{\mathbb{1}_S : S \subseteq [0,1] \text{ is a finite set}\}$, and assume that $\mathbb{P}$ has density. This function class is *not* a Glivenko-Cantelli class. First note that $\mathcal{F}_1 \subseteq \mathcal{F}_2$, so if $\mathcal{F}_2$ is GC, then $\mathcal{F}_1$ is GC. So large function classes are less likely to be GC. To show that the function class is not GC, we can find a function in the function class which makes these two quantites different. Pick $S = \{X_i : i \in [n]\}$, so

$$\sup_{S \text{ finite}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in S\}} - \mathbb{E}[\mathbb{1}_{\{X_i \in S\}}] \right| \geq |1 - 0|.$$

This lower bound holds for every n, so this difference will never go to 0.

Our next goal is to study some methods for upper/lower bounding $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$. We will see

- Rademacher complexity and VC dimension (chapter 4 of Wainwright's book),
- Metric entropy method and chaining (chapter 5 of Wainwright's book).

1.3 Rademacher complexity

Recall that the Rademacher complexity of a set $A \subseteq \mathbb{R}^n$ is

$$\mathcal{R}(A) := \mathbb{E}_{\varepsilon \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left\{ \sup_{a \in A} \langle a, \varepsilon \rangle \right\}$$

Definition 1.3. Given a function class $\mathcal{F}$ and a fixed data set $(x_i)_{i \in [n]} \subseteq \mathcal{X}$, let

$$\mathcal{F}(x_{1:n}) := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

The **Rademacher complexity** of the function class $\mathcal{F}$ and the data set $(x_i)_{i \in [n]}$ is

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) := \mathbb{E}_{\varepsilon \overset{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right\}.$$

If we write $\mathcal{A} = \pm\mathcal{F}(x_{1:n})/n$, then we can relate Rademacher complexity of sets and function classes by

$$\widetilde{\mathcal{R}}(A) = \mathcal{R}(\mathcal{F}(x_{1:n})/n),$$

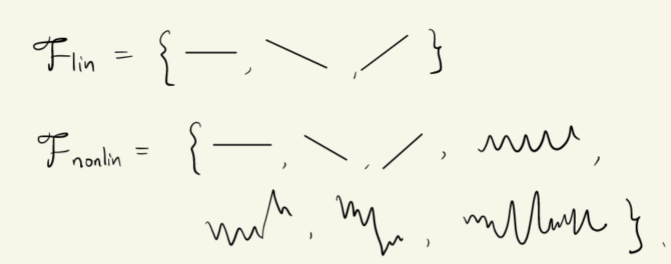
where $\widetilde{\mathcal{R}}$ denotes the Rademacher complexity of a set.

Definition 1.4. Given a function class $\mathcal{F}$ and a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, let $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathbb{P}$. The **Rademacher complexity** of the function class $\mathcal{F}$ is

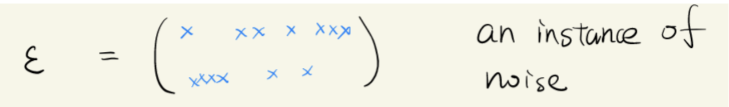
$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_{X_i \overset{\text{iid}}{\sim} \mathbb{P}}[\mathcal{R}(\mathcal{F}(X_{1:n})/n)].$$

First, observe that if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $\mathcal{R}_n(\mathcal{F}_1) \leq \mathcal{R}_n(\mathcal{F}_2)$, so this is a measure of the size of a function class.

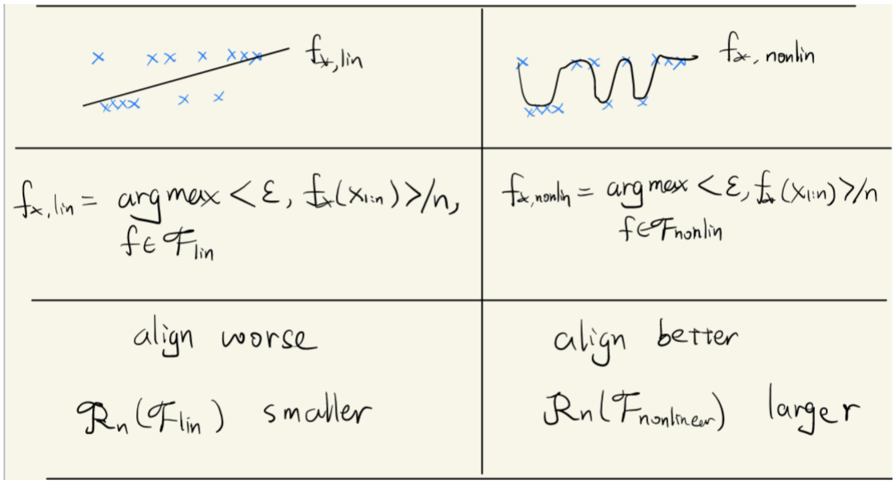
Example 1.4. Consider comparing two function classes:


$\mathcal{F}_{\text{lin}} = \{$
$\}$
$\mathcal{F}_{\text{nonlin}} = \{$
$\}$

The notion of Rademacher complexity measures how well functions in the function class can align with Rademacher noise.



Here is the picture of what the comparison would look like:


$f_{*,\text{lin}}$
$f_{*,\text{nonlin}}$
$f_{*,\text{lin}} = \arg\max_{f\in\mathcal{F}_{\text{lin}}} \langle \varepsilon, f(x_{1:n})\rangle/n,$
$f_{*,\text{nonlin}} = \arg\max_{f\in\mathcal{F}_{\text{nonlin}}} \langle \varepsilon, f(x_{1:n})\rangle/n$
align worse
$\mathcal{R}_n(\mathcal{F}_{\text{lin}})$ smaller
align better
$\mathcal{R}_n(\mathcal{F}_{\text{nonlinear}})$ larger

Example 1.5. Let $\psi : \mathbb{R}^d \to \mathbb{R}^p$ be a fixed feature map, and consider the function class

$$\mathcal{F} = \{f(x) = \langle \psi(x), \theta \rangle : \|\theta\|_2 \leq B\}.$$

Then the Rachemacher complexity of this function class is

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{\|\theta\|_2\leq B}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \psi(X_i), \theta\rangle\right|\right] \\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\sup_{\|\theta\|_2\leq B}\left|\varepsilon_i \langle \frac{1}{n}\sum_{i=1}^n \psi(X_i), \theta\rangle\right|\right] \\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi(X_i)\right\|_2\right] \cdot B
\end{aligned}$$

Using Cauchy-Schwarz,

$$\begin{aligned}
&\leq \mathbb{E}_{X_i,\varepsilon_i}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \psi(X_i)\right\|_2^2\right]^{1/2} \cdot B \\
&= \mathbb{E}_{X_i,\varepsilon_i}\left[\frac{1}{n^2}\sum_{i=1}^n \varepsilon_i^2 \|\psi(X_i)\|_2^2\right]^{1/2} \cdot B \\
&= \frac{B}{\sqrt{n}}\mathbb{E}[\|\psi(X)\|_2^2]^{1/2}.
\end{aligned}$$

Why introduce Rademacher complexity?

1. We will show that
$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \approx \mathcal{R}_n(\mathcal{F}).$$

2. The Rademacher complexity is easier to upper bound. We will have tools to upper bound it, such as
   - contraction inequality,
   - VC dimension,
   - fat-shattering dimension.

1.4 An upper bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ via $\mathcal{R}_n(\mathcal{F})$

Proposition 1.1. *For any function class $\mathcal{F}$ and distribution $\mathbb{P}$,*

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof. Let $Y_i \overset{\text{iid}}{\sim} X_i$ be independent of X_i. Then

$$\begin{aligned}
\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)]\right|\right] \\
&= \mathbb{E}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{Y_{1:n}}[f(Y_i)]\right|\right] \\
&\leq \mathbb{E}_{X_{1:n},Y_{1:n}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - f(Y_i))\right|\right]
\end{aligned}$$

We can introduce a Rademacher random variable without changing the distribution.

$$\begin{aligned}
&= \mathbb{E}_{X_{1:n},Y_{1:n},\varepsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i) - f(Y_i))\right|\right] \\
&\leq \mathbb{E}_{X_{1:n},Y_{1:n},\varepsilon_{1:n}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| - \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i))\right|\right] \\
&\leq 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}$$

□

Next lecture, we will use a similar argument to show that if $\overline{\mathcal{F}} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$, then

$$\mathcal{R}_n(\overline{\mathcal{F}}) \leq 2\,\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}].$$