SDS 384 11: Theoretical Statistics
Lecture 11: Uniform Law of Large Numbers

Purnamrita Sarkar
Department of Statistics and Data Science
The University of Texas at Austin
Uniform convergence of CDFs

• Given $X_1, \ldots, X_n \overset{iid}{\sim} F$, where $F$ is the CDF of some unknown density.
• A natural estimate of $F$ is given by

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

• $1_{-\infty, t}$ is the indicator function for $\{x \leq t\}$
• $\hat{F}_n(t)$ is the empirical CDF.
• Note that this is unbiased since $E[\hat{F}_n(t)] = F(t)$
Law of large numbers

• For any fixed $t \in \mathbb{R}$, LLN states that $\hat{F}_n(t) \xrightarrow{P} F(t)$

![Graph showing population and empirical CDFs](image)

**Figure 4-1.** Plots of population and empirical CDF functions for the uniform distribution on $[0, 1]$. (a) Empirical CDF based on $n = 10$ samples. (b) Empirical CDF based on $n = 100$ samples.

[Taken from Martin Wainwright’s book]
Why the empirical CDF?

- A statistical functional maps a CDF to a real number.
- Say you want to estimate a statistical functional $\gamma(F)$.
- A natural estimator uses the “plug in” principle, i.e. $\gamma(\hat{F}_n)$.
- Understanding the properties of the empirical CDF will help us understand why this plug in estimator is a good estimator.
Examples of functionals-expectation

Example

Given some integrable function \( g \), the expectation functional is given by

\[
\gamma_g(F) := \int g(x) dF(x)
\]

- Let \( g(x) := x \)
- \( \gamma_g(F) = E[X] \)
- \( \gamma_g(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \), which is the sample average.
- For general \( g \), \( \gamma_g(\hat{F}_n) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \)
**Examples of functionals—quantile**

**Example**

Given some \( \alpha \in [0, 1] \), the quantile functional \( Q_\alpha \) is given by

\[
Q_\alpha(F) := \inf\{ t \in \mathbb{R} | F(t) \geq \alpha \}
\]

- The median corresponds to the special case \( \alpha = 1/2 \)
- The plug in estimator is given by the sample quantile.

\[
Q_\alpha(\hat{F}_n) = \inf\{ t \in \mathbb{R} | \hat{F}_n(t) \geq \alpha \}.
\]

- The question is whether the estimate converges in some sense to the truth.
  - Note that the above function is nonlinear and so we cannot use law of large numbers to show consistency.
How do we measure consistency?

- First define $\|F - G\|_\infty := \sup_{t \in \mathbb{R}} |G(t) - F(t)|$ to measure the distance between two CDF’s $F$ and $G$.
- Now define continuity of a functional w.r.t this norm.
- We will say that $\gamma$ is continuous at $F$ in the sup-norm if

$$\forall \epsilon > 0, \exists \delta > 0, \text{ s.t. } \|G - F\|_\infty \leq \delta \Rightarrow |\gamma(G) - \gamma(F)| \leq \epsilon.$$  
- This essentially means that in order to show consistency of a plug-in estimator we need to show that $\|\hat{F}_n - F\|_\infty$ converges to zero.
The Glivenko Cantelli theorem

Theorem
For any distribution the empirical CDF $\hat{F}_n$ is a strongly consistent estimator of the population CDF $F$ in the uniform norm, i.e.

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

- We prove this later.
• Consider the function class $\mathcal{F}$ of integrable real-valued functions.

• Let $\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i} f(X_i) - E[f] \right|

**Definition**

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

• Can also be defined in a stronger sense.

• We say that $\mathcal{F}$ satisfies the strong **Glivenko-Cantelli** law if the above quantity converges to zero a.s.
The classical Glivenko Cantelli theorem

- Consider the function class $\mathcal{F}$ of indicator functions of the form $\mathcal{F} := \{I_{(-\infty, t]}(.) | t \in \mathbb{R}\}$.
- For a fixed $t \in \mathbb{R}$, $E[I_{(-\infty, t]}(X)] = P(X \leq t) = F(t)$
- So the classical GC theorem corresponds to a strong uniform law for the above class.
Example

Let $S$ be the class of all subsets of $[0,1]$ such that the subset $S$ has a finite number of elements. Now consider $\mathcal{F}_S := \{1_S(.)| S \in S\}$. Let $X_i \overset{iid}{\sim} P$ s.t. $P$ is a distribution over $[0,1]$ and $P$ has no atoms, i.e. $P(\{x\}) = 0, \forall x \in [0,1]$. This class is not a GC class for $P$.

- First note that $P[S] = 0, \forall S \in S$.
- Let $X = \{X_1, \ldots, X_n\}$
- We see that $X \in S$, and $P_n[X] = 1$.
- $\sup_{S \in S} |P_n[S] - P[S]| = 1 - 0 = 1$
Coming back to functionals

- We saw that functionals help us look at quantities like quantiles, means, etc. But is that all?
- As it turns out they help enormously for empirical risk minimization too.
- Consider the indexed family of probability distributions
  \[ \mathcal{P}_\Theta := \{P_\theta | \theta \in \Theta \} \]
- Let \( X = \{X_1, \ldots, X_n\} \overset{iid}{\sim} P_{\theta^*}, \) where \( \theta^* \in \Theta \)
- This \( \theta^* \) could lie in some \( d \) dimensional space
  - Take for example the problem of estimating the means of a Mixture of Gaussians.
- This \( \theta^* \) could also be lying in some function class, which will give us a non-parametric estimation problem.
Estimating the true $\theta^*$

- In these cases, we estimate $\theta^*$ by minimizing a loss function of the form $\mathcal{L}_\theta(x)$ which measures how well $P_\theta$ represents or fits the unknown distributions.

- Empirical risk minimization is based on the objective function, also known as the empirical risk

$$\hat{R}_n(\theta, \theta^*) = \frac{1}{n} \sum_i \mathcal{L}_\theta(X_i)$$

- The population risk is given by

$$R(\theta, \theta^*) := \mathbb{E}_{\theta^*} \left[ \mathcal{L}_\theta(X_1) \right]_{E_{X_1 \sim P_{\theta^*}}}$$
Empirical risk minimization

- Sometimes, we minimize empirical risk over some subset $\Theta_0 \in \Theta$, to get $\hat{\theta}$
- The statistical question is how small is the excess risk
  
  \[
  R(\hat{\theta}, \theta^*) - \inf_{\theta \in \Theta_0} R(\theta, \theta^*)
  \]
- Now we will look at some examples
Example: Maximum Likelihood

Consider a family of distributions \( \{P_\theta, \theta \in \Theta\} \), each with a strictly positive density \( p_\theta \). Now suppose that we are given \( X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*} \).

We would like to estimate the unknown parameter \( \theta \). In order to do so, we consider the objective function

\[
\mathcal{L}_\theta(x) := \log \frac{p_{\theta^*}(x)}{p_\theta(x)}
\]

- The maximum likelihood estimate is indeed

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_i \mathcal{L}_\theta(X_i).
\]

- The population risk is \( R(\theta, \theta^*) = E_{\theta^*} \log \frac{p_{\theta^*}(X)}{p_\theta(X)} \), which is the KL divergence between the fitted and true densities.
Example: binary classification

Example

You observe $n$ i.i.d samples $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$, where $X_i$ is a set of $d$ features, and $Y_i$ corresponds to the label. One can assume that $X_i \sim P_X$ and $Y_i \sim P_{Y|X=X_i}$. In this context we want to estimate some function $f : \mathbb{R}^d \rightarrow \{-1, 1\}$ which minimizes the probability of misclassification. We use

$$
\mathcal{L}_f(x, y) := \begin{cases} 
1 & \text{if } f(x) = y \\
0 & \text{otherwise}
\end{cases}
$$
Empirical risk minimization

- For equally probable classes, the Bayes classifier $f^\ast$ is given by:

$$f^\ast(x) := \begin{cases} 
1 & \text{if } P(Y = 1|X = x) \geq P(Y = -1|X = x) \\
-1 & \text{if } P(Y = 1|X = x) < P(Y = -1|X = x)
\end{cases}$$

- In practice, since the odds ratio is unknown, we often minimize:

$$\hat{R}_n(f, f^\ast) = \sum_{i=1}^{n} 1_{f(X_i) \neq Y_i}.$$ 

- The above is also the training error rate.

- Typically we minimize the above over some restricted set of decision rules.
Empirical risk minimization

- Our goal is to understand the behavior of the excess risk.
- Recall that we want to bound $R(\hat{\theta}, \theta^*) - \inf_{\theta \in \Theta_0} R(\theta, \theta^*)$, aka, $\delta R(\hat{\theta}, \theta^*)$.
- Assume for convenience that the infimum over $\theta \in \Theta_0$ is achieved at $\theta_0 \in \Theta_0$.
- $\delta R(\hat{\theta}, \theta^*)$ equals
  \[
  \underbrace{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)}_T + \underbrace{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)}_T + \underbrace{\hat{R}_n(\theta_0, \theta^*) - R(\theta_0, \theta^*)}_T
  \]

- $T_3$ is just the deviation of a sum of bounded and iid random variables from its expectation. So this can be easily bounded using tools like Hoeffding etc.
Empirical risk minimization

- \( T_3 = \frac{1}{n} \sum_i \mathcal{L}_{\theta_0}(X_i) - E[\mathcal{L}_{\theta_0}(X_i)] \)

- When \( \mathcal{L} \) is a bounded loss function, we can use techniques we have learned so far.

- Let's look at \( -T_1 = \frac{1}{n} \sum_i \mathcal{L}_{\hat{\theta}}(X_i) - E[\mathcal{L}_{\hat{\theta}}(X_i)] \)

- This again is much harder to analyze since \( \hat{\theta} \) is a function of \( X_1, \ldots, X_n \).

- Typically we bound this using

  \[
  T_1 \leq \sup_{\theta \in \Theta_0} \left| \frac{1}{n} \sum_i \mathcal{L}_{\theta}(X_i) - E[\mathcal{L}_{\theta}(X_i)] \right| =: \| \hat{P}_n - P \|_{\mathcal{L}(\Theta_0)}
  \]

- Where \( \mathcal{L}(\Theta_0) \) is the loss class \( \{ \mathcal{L}_\theta | \theta \in \Theta_0 \} \)
Empirical Risk Minimization

- $T_3 = \frac{1}{n} \sum_i \mathcal{L}_{\theta_0}(X_i) - E[\mathcal{L}_{\theta_0}(X_i)] \leq \|\hat{P}_n - P\|_{\mathcal{L}(\Theta_0)}$

- $\delta R(\hat{\theta}, \theta^*) \leq 2\|\hat{P}_n - P\|_{\mathcal{L}(\Theta_0)}$

- Now we will establish an uniform law of large numbers for the loss class $\mathcal{L}(\Theta_0)$