Stat 6390.001: High Dimensional Probability and Statistics, Fall 2026 **Topic 10**

Uniform Law of Large Numbers and Empirical Risk Minimization

Lecturer: Yun Wei *Scribe: Thomas Yeboah*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Summary of Notation for Empirical Measures and Empirical Processes

Notation	Definition	Interpretation
$\mathbb{P}$	Probability measure	
$\mathbb{P}f$	$\mathbb{P}f = \int f(x)\,\mathbb{P}(dx)$	$\mathbb{E}_{\mathbb{P}}[f(X)]$
$\{\mathbb{P}f\}_{f\in\mathcal{F}}$	Functions indexed by $\mathcal{F}$	
$\mathbb{P}_n$	$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ Discrete random measure	$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_A(X_i)$ $\mathbb{P}_n(X_i) = \frac{1}{n},\ \forall i$
$\mathbb{P}_n f$	$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$	Empirical mean of $f(X)$
$\{\mathbb{P}_n f\}_{f\in\mathcal{F}}$	$\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) : f\in\mathcal{F}\right\}$	Random process
$\mathbb{P}_n f - \mathbb{P}f$	$\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{\mathbb{P}}[f(X)]$	Random variable
$\{\mathbb{P}_n f - \mathbb{P}f\}_{f\in\mathcal{F}}$	$\left\{\frac{1}{n}\sum_{i=1}^{n}\left[f(X_i) - \mathbb{E}_{\mathbb{P}}f(X)\right]\right\}_{f\in\mathcal{F}}$	Centered empirical process
$\lVert\mathbb{P}_n - \mathbb{P}\rVert_{\mathcal{F}}$	$\sup_{f\in\mathcal{F}}\lvert\mathbb{P}_n f - \mathbb{P}f\rvert$	Supremum of the empirical process

1 Glivenko-Cantelli Class

Definition 1.1 (Glivenko–Cantelli Class)**.** We say that $\mathcal{F}$ is a **Glivenko–Cantelli class** for $\mathbb{P}$ if

$$\lVert\mathbb{P}_n - \mathbb{P}\rVert_{\mathcal{F}} \xrightarrow{P} 0, \qquad n\to\infty.$$

If

$$\lVert\mathbb{P}_n - \mathbb{P}\rVert_{\mathcal{F}} \xrightarrow{\text{a.s.}} 0,$$

then $\mathcal{F}$ is called a **strong Glivenko–Cantelli class**.

Example 1.2 (A Strong Glivenko–Cantelli Class)**.** Consider the class

$$\mathcal{F} = \left\{\mathbf{1}_{(-\infty,t]}(\cdot) : t\in\mathbb{R}\right\}.$$

For

$$f_t(\cdot) = \mathbf{1}_{(-\infty,t]}(\cdot) \in \mathcal{F},$$

we have

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

Also,

$$\begin{aligned}\mathbb{P}_n f_t &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i) \\ &= \widehat{F}_n(t).\end{aligned}$$

Therefore,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\widehat{F}_n - F\|_\infty.$$

By the classical Glivenko–Cantelli theorem,

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

for every distribution function F.

Hence,

$$\boxed{\mathcal{F} \text{ is a strong Glivenko–Cantelli class.}}$$

Example 1.3 (A Class That Is Not Glivenko–Cantelli)**.** Let $\mathcal{S}$ be the class of all finite subsets A of $[0,1]$.

Define

$$\mathcal{F}_{\mathcal{S}} = \{\mathbf{1}_A(\cdot) : A \in \mathcal{S}\}.$$

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathbb{P},$$

where $\mathbb{P}$ is a probability measure on $[0,1]$ that has no atoms.

Since $\mathbb{P}$ has no atoms,

$$\mathbb{P}(\{y\}) = 0, \qquad \forall\, y \in [0,1].$$

Let

$$A_0 = \{X_1, \ldots, X_n\} \in \mathcal{S}$$

and

$$f_0 = \mathbf{1}_{A_0}(\cdot) \in \mathcal{F}_{\mathcal{S}}.$$

Then

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_{\mathcal{S}}} \geq |\mathbb{P}_n f_0 - \mathbb{P}f_0|.$$

Since $\mathbb{P}$ has no atoms,

$$\begin{aligned}\mathbb{P}f_0 &= \mathbb{P}\mathbf{1}_{A_0}(\cdot) \\ &= \mathbb{P}(A_0) \\ &= 0.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{P}_n f_0 &= \mathbb{P}_n \mathbf{1}_{A_0}(\cdot) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{A_0}(X_i) \\ &= 1.\end{aligned}$$

Therefore,

$$\boxed{\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}_{\mathcal{S}}} \geq 1 > 0.}$$

Hence, $\mathcal{F}_{\mathcal{S}}$ is not a Glivenko–Cantelli class for any atomless probability measure $\mathbb{P}$.

2 Empirical Risk Minimization

Let

$$\{\mathbb{P}_\theta : \theta \in \Omega\}$$

be a parameterized family of distributions, and suppose

$$X_i \overset{\text{iid}}{\sim} \mathbb{P}_{\theta^*}.$$

The parameter space Ω may be finite-dimensional or infinite-dimensional.
Let $L_\theta(X)$ denote the cost function.
The **population risk** is

$$R(\theta, \theta^*) = \mathbb{E}_{\theta^*}[L_\theta(X)].$$

The **empirical risk** is

$$\widehat{R}_n(\theta, \theta^*) = \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i).$$

An empirical risk minimizer is

$$\widehat{\theta}_n \in \arg\min_{\theta \in \Omega_0} \widehat{R}_n(\theta, \theta^*).$$

We are interested in the **excess population risk**

$$\boxed{\begin{aligned} E(\widehat{\theta}_n, \theta^*) = R(\widehat{\theta}_n, \theta^*) \\ - \inf_{\theta \in \Omega_0} R(\theta, \theta^*). \end{aligned}}$$

Example 2.1 (Maximum Likelihood Estimation)**.** Let

$$\{P_\theta : \theta \in \Omega\}$$

be a family of probability distributions with strictly positive densities p_θ with respect to a common underlying measure.

Define

$$L_\theta(x) = \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right).$$

Then

$$\begin{aligned} R(\theta, \theta^*) &= \mathbb{E}_{\theta^*}[L_\theta(X)] \\ &= \int p_{\theta^*}(x) \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right) dx \\ &= D(P_{\theta^*}, P_\theta). \end{aligned}$$

Thus, under identifiability,

$$\theta^* = \arg\min_{\theta \in \Omega_0} R(\theta, \theta^*).$$

The empirical risk is

$$\begin{aligned} \widehat{R}_n(\theta, \theta^*) &= \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right). \end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{\theta}_n &\in \arg\min_{\theta\in\Omega_0} \widehat{R}_n(\theta,\theta^*)\\ &= \arg\min_{\theta\in\Omega_0}\left\{\frac{1}{n}\sum_{i=1}^n \log p_{\theta^*}(X_i) - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)\right\}.\end{aligned}$$

Since

$$\frac{1}{n}\sum_{i=1}^n \log p_{\theta^*}(X_i)$$

does not depend on θ, we obtain

$$\boxed{\widehat{\theta}_n \in \arg\max_{\theta\in\Omega_0} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i).}$$

This is the **maximum likelihood estimator**. If $\theta^* \in \Omega_0$,

$$\begin{aligned}E(\widehat{\theta}_n,\theta^*) &= R(\widehat{\theta}_n,\theta^*) - \inf_{\theta\in\Omega_0} R(\theta,\theta^*)\\ &= D(P_{\theta^*}, P_{\widehat{\theta}_n}) - 0\\ &= D(P_{\theta^*}, P_{\widehat{\theta}_n}).\end{aligned}$$

Example 2.2 (Binary Classification)**.** Suppose

$$(X_i,Y_i) \in \mathbb{R}^d \times \{-1,+1\}, \qquad i=1,\dots,n,$$

and

$$(X_i,Y_i) \overset{\text{iid}}{\sim} \mathbb{P}_{X,Y}.$$

In particular,

$$X_i \overset{\text{iid}}{\sim} \mathbb{P}_X, \qquad Y_i \mid X_i \sim \mathbb{P}_{Y|X}.$$

Define

$$\psi(x) = \frac{\mathbb{P}(Y=+1 \mid X=x)}{\mathbb{P}(Y=-1 \mid X=x)}.$$

We want to find a classifier

$$f : \mathbb{R}^d \longrightarrow \{-1,+1\}$$

that minimizes

$$\mathbb{P}_{X,Y}\big(f(X)\neq Y\big).$$

Define the loss function

$$L_f(x,y) := \mathbf{1}\{f(x)\neq y\}.$$

Then the risk of a classifier f is

$$\begin{aligned}R(f,f^*) &= \mathbb{E}_{X,Y}\mathbf{1}\{f(X)\neq Y\}\\ &= \mathbb{P}_{X,Y}\big(f(X)\neq Y\big).\end{aligned}$$

The **Bayes classifier** is

$$\boxed{f^*(x) = \begin{cases}+1, & \text{if } \psi(x)\geq 1,\\ -1, & \text{if } \psi(x)<1.\end{cases}}$$

The empirical risk is

$$\widehat{R}_n(f, f^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{f(X_i) \neq Y_i\}.$$

Also,

$$1 - \mathbb{P}_{X,Y}\big(f(X) \neq Y\big) = \mathbb{P}_{X,Y}\big(f(X) = Y\big).$$

By conditioning on X,

$$\begin{aligned}
&\mathbb{P}_{X,Y}\big(f(X) = Y\big) \\
&\quad = \mathbb{E}_X\left[\mathbb{P}_{Y|X}\big(f(X) = Y \mid X\big)\right] \\
&\quad = \mathbb{E}_X\Big[\mathbb{P}(Y = +1 \mid X)\mathbf{1}\{f(X) = +1\} \\
&\qquad\qquad + \mathbb{P}(Y = -1 \mid X)\mathbf{1}\{f(X) = -1\}\Big].
\end{aligned}$$

Thus the Bayes classifier f^* maximizes $\mathbb{P}_{X,Y}\big(f(X) = Y\big)$.

References

[1] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.