I will try to give a high level overview of what we plan to cover in this class. The tentative list of topics to be covered is:


1 Empirical Process Theory

1.1 Motivation via Goodness of Fit Testing

The problem of testing goodness of fit provides a good classical motivation for the study of empirical processes.

Consider the statistical problem of goodness of fit hypothesis testing where one observes an i.i.d sample $X_1, \ldots, X_n$ from a distribution $F$ and wants to test the null hypothesis $H_0 : F = F_0$ against the alternative hypothesis $H_1 : F \neq F_0$. Here $F_0$ is a fixed distribution function.

Karl Pearson provided the following solution for this problem:

1. Partition the real line into $k$ disjoint sets: $C_1, \ldots, C_k$. For each $i = 1, \ldots, k$, let $Y_i$ denote the number of observations among $X_1, \ldots, X_n$ that lie in $C_i$. Also let $p_i := \Pr\{X_1 \in C_i\}$.

2. Calculate the quantity:

$$X^2 := \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}.$$ 

Under the null hypothesis (assuming that $p_i > 0$ for each $i$), $X^2$ converges in distribution to a chi-squared distribution with $k - 1$ degrees of freedom. Use this to obtain a $p$-value for the test.

To prove that $X^2$ converges to $\chi^2_{k-1}$, observe first that by the multivariate CLT, $Z_{k-1} := ((Y_1 - np_1)/\sqrt{n}, \ldots, (Y_{k-1} - np_{k-1})/\sqrt{n})^T$ converges in distribution to $N_{k-1}(0, \Sigma)$ where the entries of $\Sigma$ are given by $\sigma_{ii} = p_i(1 - p_i)$ and $\sigma_{ij} = -p_ip_j$ for $i \neq j$. Therefore, $Z_{k-1}^T \Sigma^{-1} Z_{k-1}$ converges in distribution to $\chi^2_{k-1}$. Use the Sherman-Morrison formula to show that $\Sigma^{-1} := diag(p_1^{-1}, \ldots, p_{k-1}^{-1}) - 11^T p_k^{-1}$. This immediately gives that $X^2 = Z_{k-1}^T \Sigma^{-1} Z_{k-1}$ which completes the proof.
This is called the chi-squared goodness of fit test. It is widely used but has the following important shortcoming. Only considering the probabilities of \( F_0 \) in the cells \( C_1, \ldots, C_k \) results in a loss of information. In other words, the test will never reject the null hypothesis \( H_0 \) if the data come from a distribution which gives probabilities to \( C_1, \ldots, C_k \) that are nearly equal to those given by \( F_0 \). Therefore, the chi-squared test has low power against such alternatives.

More satisfactory goodness of fit tests were proposed by Cramer, Von Mises, Kolmogorov and Smirnov. These tests are based on various divergences between the hypothesized distribution function \( F_0 \) and the empirical distribution function \( F_n \). The empirical distribution function \( F_n \) is defined as

\[
F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \{X_i \leq x\}.
\]

In other words, for each \( x \in \mathbb{R} \), the quantity \( nF_n(x) \) simply counts the number of \( X_i \) that are less than or equal to \( x \).

For testing \( H_0 : F = F_0 \), Kolmogorov recommended working with the quantity

\[
D_n := \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|
\]

and rejecting \( H_0 \) when \( D_n \) is large. To calculate the \( p \)-value of this test, the null distribution (i.e., the distribution of \( D_n \) under \( H_0 \)) needs to be determined. An interesting property about the null distribution of \( F_0 \) is that the null distribution is the same whenever \( F_0 \) is continuous:

**Lemma 1.1.** The null distribution of \( D_n \) i.e., the distribution of \( D_n \) when \( X_1, \ldots, X_n \) are i.i.d \( F_0 \) is the same for all continuous distributions \( F_0 \).

**Proof.** Exercise. Use the quantile transformation.

Because of the above lemma, we can compute the null distribution of \( D_n \) assuming that \( F_0 \) is the distribution function of a uniformly distributed random variable on \((0, 1)\). In other words, the null distribution of \( D_n \) is the same as that of \( \sup_{t \in (0,1)} |U_n(t)| \) where

\[
U_n(t) := \sqrt{n} (G_n(t) - t) \quad \text{with} \quad G_n(t) := \frac{1}{n} \sum_{i=1}^{n} \{\xi_i \leq t\}
\]

and \( \xi_1, \ldots, \xi_n \) are i.i.d \( \text{unif}(0,1) \) random variables. The function \( t \to U_n(t) \) is called the uniform empirical process.

Kolmogorov proved that

\[
\lim_{n \to \infty} \mathbb{P}_{H_0} \{D_n \leq x\} = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2} \quad \text{for every } x > 0. \tag{1}
\]

This formula can be used to calculate approximate \( p \)-values for this test.

The Kolmogorov test for \( H_0 : F = F_0 \) is only one of a large class of tests that are based on some measure of distance between the empirical cdf \( F_n \) and \( F_0 \). Two other such tests are based on the following quantities:

1. **Cramer - Von Mises Statistic**: Defined by

\[
W_n := n \int (F_n(x) - F_0(x))^2 dF_0(x)
\]
2. **Anderson-Darling Statistic**: Defined by

\[ A_n := n \int \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x). \]

Smirnov also considered the following statistics in the context of certain other tests:

\[ D^+_n := \sqrt{n} \sup_x (F_n(x) - F_0(x)) \quad \text{and} \quad D^-_n := \sqrt{n} \sup_x (F_0(x) - F_n(x)). \]

All of these quantities have the property that their null distribution (i.e., when \( F = F_0 \)) is the same for all continuous \( F_0 \). Thus one may assume that \( F_0 \) is the uniform distribution for computing their null distribution. And in this case, all these quantities can be written in terms of the uniform empirical process \( U_n(t) \).

Initially, the asymptotic distributions of these quantities were determined on a case by case basis without a unified technique. Doob realized that it should be possible to obtain these distributions using some basic properties of the uniform empirical process. By the multivariate Central Limit Theorem, for every \( k \geq 1 \) and \( 0 < t_1, \ldots, t_k < 1 \), the random vector \((U_n(t_1), \ldots, U_n(t_k))\) converges in distribution to \( N_k(0, \Sigma) \) where \( \Sigma(i, j) := t_i \wedge t_j - t_i t_j \) (here \( a \wedge b := \min(a, b) \)). This limiting distribution \( N(0, \Sigma) \) is the same as the distribution of \((U(t_1), \ldots, U(t_k))\) where \( U(t) \) is the Brownian Bridge characterized by the following two properties:

1. \( U(0) = U(1) = 0 \). For every \( t \in (0, 1) \), \( U(t) \) is a random variable.
2. For every \( k \geq 1 \) and \( t_1, \ldots, t_k \in (0, 1) \), the random vector \((U(t_1), \ldots, U(t_k))\) has the \( N(0, \Sigma) \) distribution.
3. The function \( t \to U(t) \) is (almost surely) continuous on \([0, 1]\).

Doob therefore conjectured that the uniform empirical process \( \{U_n(t) : t \in [0, 1]\} \) must converge in some sense to a Brownian Bridge \( \{U(t) : t \in [0, 1]\} \). Hopefully, this notion of convergence will be strong enough to yield that various functionals of \( U_n(t) \) will converge to the corresponding functionals of \( U(t) \). Motivated by this idea, Doob calculated the distribution of \( \sup_{t \in [0, 1]} |U(t)| \) and proved that

\[ \mathbb{P} \left\{ \sup_{t \in [0, 1]} |U(t)| \leq x \right\} = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 x^2} \quad \text{for every } x > 0. \]

If the convergence of \( U_n(t) \) to \( U(t) \) is made rigorous, this would provide an alternative proof of Kolmogorov’s result (1). In addition, this would allow for the derivation of asymptotic distributions of \( W_n, A_n, D^+_n \) and \( D^-_n \) as well.

Donsker accomplished this by first establishing a rigorous theory of convergence of stochastic processes and then proving that the uniform empirical process converges to Brownian Bridge. Donsker’s theorem is one of two basic facts about the uniform empirical process. The second basic fact is the Glivenko-Cantelli Theorem which states that \( \sup_{0 \leq t \leq 1} |G_n(t) - t| \) converges to 0 almost surely.

We will study these two theorems (Glivenko-Cantelli and Donsker) and their extensions to more general empirical processes (which are defined next).

### 1.2 Empirical Processes

An empirical process is a special kind of a stochastic process (a stochastic process is just a collection of random variables). For a rigorous definition, we need the following basic setting.
Let \( \xi_1, \xi_2, \ldots \) denote a sequence of independent and identically distributed random elements taking values in an abstract measurable space \( \mathcal{X} \). This means that each \( \xi_i \) is a measurable map from a probability space \((\Omega, \mathcal{A}, \mathbb{P})\) to \( \mathcal{X} \). Let \( P \) denote the distribution of \( \xi_1 \) i.e., \( P \) is the probability measure on \( \mathcal{X} \) defined by \( P(A) := \mathbb{P}\{\xi_1 \in A\} \).

Let \( \mathcal{F} \) denote a class of measurable real-valued functions on \( \mathcal{X} \).

For each \( n \geq 1 \), let \( P_n \) (called the empirical measure) denote the random discrete probability measure which puts mass \( 1/n \) at each of the \( n \) points \( \xi_1, \ldots, \xi_n \). For each real valued function \( f \) on \( \mathcal{X} \), define

\[
P f := \int_{\mathcal{X}} f dP = \mathbb{E}(f(\xi_1)) \quad \text{and} \quad P_n f := \int_{\mathcal{X}} f dP_n = \frac{1}{n} \sum_{i=1}^{n} f(\xi_i).
\]

Let us assume that \( Pf \) exists for each \( f \in \mathcal{F} \). The goal of empirical process theory is to study the properties of the approximation of \( Pf \) by \( P_n f \) uniformly in \( \mathcal{F} \).

The collection of random variables \( \nu_n(f) := \sqrt{n}(P_n f - Pf) \) as \( f \) varies over \( \mathcal{F} \) is called the Empirical Process.

**Example 1.2 (Uniform empirical process).** Suppose \( \mathcal{X} = [0, 1] \) and \( P \) is the uniform probability measure on \( [0, 1] \). Also let \( \mathcal{F} := \{I([0,t]) : t \in [0,1]\} \). Functions in \( \mathcal{F} \) can then simply be indexed by \( t \) and then it is easy to see that \( \nu_n(f), f \in \mathcal{F} \) is simply the uniform empirical process \( U_n(t), t \in [0,1] \).

Mostly, we will focus on the following two types of results:

1. **Glivenko-Cantelli:** Under what conditions on \( \mathcal{F} \) does \( \sup_{f \in \mathcal{F}} |P_n f - Pf| \) converge to zero almost surely?
2. **Donsker:** Under what conditions on \( \mathcal{F} \) does \( \nu_n(f), f \in \mathcal{F} \) converge as a process to some limit object as \( n \to \infty \).

I will follow Pollard’s book on convergence in stochastic processes for this stuff.

## 2 Asymptotic theory of M-estimation

Empirical process results play a key role in the theory of M-estimators. M-estimators are defined as maximizers of sums of functions of the data. Specifically, consider independent data \( X_1, \ldots, X_n \) taking values in \( \mathcal{X} \) and a class of functions \( \mathcal{F} \) on \( \mathcal{X} \). Then \( \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) \) is an M-estimator.

Examples of M-estimators are:

1. Maximum Likelihood Estimators: if we have i.i.d data and a parametric model \( p_\theta, \theta \in \Theta \) for the data \( X_1, \ldots, X_n \) and we take \( \mathcal{F} := \{ \log p_\theta : \theta \in \Theta \} \), then we get the MLE.
2. Least squares estimators: Suppose that the data is of the form \( (X_i, Y_i) \) for \( i = 1, \ldots, n \) and \( \mathcal{F} := \{(y - f_\theta(x))^2 : \theta \in \Theta\} \), then the relevant M-estimator is the least squares estimator over \( \Theta \).
3. Robust regression estimators: Again consider data of the form \( (X_i, Y_i) \) for \( i = 1, \ldots, n \) with and \( \mathcal{F} := \{m(y - \theta^T x) : \theta \in \mathbb{R}^d\} \). Choices for \( m \) include \( m(x) := |x| \), Huber functions etc.

Closely related to the M-estimators are the Z-estimators, which are defined as solutions to a system of equations of the form \( \sum_{i=1}^{n} \psi(X_i) = 0 \) for \( \psi \in \Psi \).

We will learn how to establish consistency, rates of convergence and the limiting distribution for M and Z-estimators. I will follow Van der Vaart and Wellner’s book on Weak convergence and empirical processes (chapters 3.1 to 3.4) for this part.
Nonasymptotic Theory of Penalized Empirical Risk Minimization

Penalized ERM can be formulated in the following way: Consider a space $\mathcal{X}$ and a probability measure $P$ on it. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. The problem is to find a function $f_{\min}$ which minimizes $Pf$ over $f \in F$ (assuming, of course, that a minimizer exists; if not, the goal becomes to find an approximate minimizer).

The probability measure $P$ is unknown however and, instead, we are given i.i.d data $X_1, \ldots, X_n$ having distribution $P$. A natural idea to solve the problem then is to consider the Penalized Empirical Risk Minimizer $\hat{f}_n$ defined by

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \{P_n f + \text{pen}(n; f)\}$$

where $\text{pen}(n; f)$ is a penalty function. The idea behind using the penalty function is to avoid overfitting.

The quality of $\hat{f}_n$ is measured by its excess risk defined as

$$\mathcal{E}(\hat{f}_n) := P \hat{f}_n - \inf_{f \in \mathcal{F}} Pf = P \hat{f}_n - Pf_{\min}.$$ 

We want nonasymptotic high probability upper bounds for $\mathcal{E}(\hat{f}_n)$. Such bounds have close connections to the so-called oracle inequalities and play a big role in model selection. I will go over some of these bounds and the main reference will be Koltchinskii’s book on *Oracle inequalities in Empirical Risk Minimization and Sparse Recovery Problems*.

Nonasymptotic deviation inequalities for suprema of empirical processes play a big role in proving bounds for $\mathcal{E}(\hat{f}_n)$. For example, consider the case when there is no penalty i.e., we want an upper bound for $\hat{\delta} := \mathcal{E}(\hat{f}_n)$ when $\hat{f}_n := \arg\min_{f \in \mathcal{F}} P_n f$. Write

$$\hat{\delta} := P \hat{f}_n - Pf_{\min} = P_n(\hat{f}_n - f_{\min}) + (P - P_n)(\hat{f}_n - f_{\min}) \leq (P - P_n)(\hat{f}_n - f_{\min}).$$

Let $\mathcal{F}(\delta) := \{f \in \mathcal{F} : \mathcal{E}(f) \leq \delta\}$. Then it follows from the above that

$$\hat{\delta} \leq \sup_{f,g \in \mathcal{F}(\delta)} |(P - P_n)(f - g)|$$

Therefore, if we have a good non-random upper bound $U_n(\delta)$ for $\sup_{f,g \in \mathcal{F}(\delta)} |(P - P_n)(f - g)|$, then we can bound $\hat{\delta}$ by the largest solution to $\delta \leq U_n(\delta)$.

A good reference for bounds on the suprema of empirical processes is the recent book by Boucheron, Lugosi and Massart on *concentration inequalities*. 