With Glivenko-Cantelli theorem, the uniform LLN is guaranteed through the convergence of Kolmogorov statistics. With Donsker theorem, the uniform CLT is guaranteed and thus we know the asymptotic behaviour of Kolmogorov statistics. In the previous discussion, we consider the setting that the underlying distribution is known. However, in real life, this is not the case! What we have is the empirical distribution and some assumption about the random source. And our goal is to find out the underlying distribution. With the help of Donsker theorem, we can construct a confidence interval of Kolmogorov statistics and draw inference about the empirical process. And finally test whether our assumption on the true distribution is correct with high probability.

0.1 Framework

Kolmogorov-Smirov test is a famous non-parametric goodness of fitting test. The test consider the Kolmogorov statistics: $D_n = \sup_{x \in \mathcal{R}} |\hat{F}_n(x) - F(x)|$, which is a *distribution-free* statistics. The convergence of D_n provides us a way to see that whether a source is sampled from the guessing distribution or not. Moreover, since the probability distribution of D_n will converge to that of a Brownian Bridge, the confidence interval can be calculated.

0.2 Convergence of Kolmogorov Statistics

Donsker theorem says that the Kolmogorov statistics will converge in distribution to the spremum of a Brownian bridge. To prove this, we prove a stronger result:

Theorem 1. *Let E_n be an empirical process with n samples and $\mathbf{B} = \{B_t : t \in [0,1]\}$ be a Brownian bridge. Then for all $t \in [0,1]$, $E_n(t) \to B_t$ almost surely as $n \to \infty$.*

To prove this theorem, we need three steps:

1. E_n converge to **B** almost surely on **finite** many of points in [0,1] as $n \to \infty$.
2. E_n with **finite** many of points in [0,1] will define the original empirical process as the number of points grows to infinity.
3. **B** with **finite** many of points in [0,1] will define the original Brownian bridge as the number of points grows to infinity.

For 2 and 3, Daniell-Kolmogorov extension theorem guaranteed the convergence. Since this is out of the discussion of this project, so we will not prove it here. However, if you are interested with it, you can find lots of resource on the internet.

As a result, our goal here is to show that the empirical process E_n will converge to the Brownian bridge $\mathbf{B}$ almost surely on **finite** many of points in [0,1] as $n \to \infty$.

Proof. First, recall the *Probability integral transform theorem* that the distribution of the CDF(cumulative distribution function) transform of any random variable is uniform on [0,1]. As a result, we can consider the empirical process of uniform distribution case first. As to the empirical process of general distribution function F, we can simply compose the $F(x)$ term to the result of uniform case. Details will be explained later.

So let's consider n uniform *i.i.d.* random variables $X_1, X_2, ..., X_n$ from $F(x) = x$ on [0,1] and use them to construct an empirical distribution $\hat{F}_n$. Now take arbitrary k points $x_1, x_2, ..., x_k$ in [0,1]. We observe the behaviour on these k finite points. Represent them in a random vector:

$$\sqrt{n}\begin{bmatrix} \hat{F}_n(x_1) - F(x_1) \\ \vdots \\ \hat{F}_n(x_k) - F(x_k) \end{bmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} \mathbf{I}\{X_i \le x_1\} - F(x_1) \\ \vdots \\ \mathbf{I}\{X_i \le x_k\} - F(x_k) \end{bmatrix} := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{Z}_i$$

where $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_n$ are *i.i.d.* k-dimensional random vectors with $E[\mathbf{Z}_i] = \mathbf{0}$ and $Cov[\mathbf{Z_i}] = \mathbf{Q}$ $\forall i$, and $\mathbf{Q}$ is the covariance matrix such that

$$\mathbf{Q}_{i,i} = F(x_i)[1 - F(x_i)]$$

and

$$\mathbf{Q}_{i,j} = F(x_i \wedge x_j) - F(x_i)F(x_j) = x_i \wedge x_j - x_i x_j$$

Note that the random part is in the random variables $X_1, ..., X_n$, that is, the empirical distribution $\hat{F}_n$. Not the arbitrary points $x_1, x_2, ..., x_k$.

Then, by the multinomial central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{Z}_i$ converges in distribution to $N_k(\mathbf{0}, \mathbf{Q})$. In other words,

$$\begin{bmatrix} E_n(x_1) \\ \vdots \\ E_n(x_k) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} W(x_1) \\ \vdots \\ W(x_k) \end{bmatrix}$$

where $\mathbf{W} = (W_1, W_2, ..., W_k)' \sim N_k(\mathbf{0}, \mathbf{Q})$.

Also, by the *Uniqueness Property* of Gaussian process and the property of $\mathbf{Q}$, we can know that $W(x_i) = B(x_i)$ $\forall i$, where B is a Brownian bridge. That is, the empirical process will converge to a Brownian bridge at finite points $x_1, x_2, ..., x_k$. Thus, 1 has been proved!

To sum up, the empirical process of uniform distribution will converge to the Brownian bridge in finite many points. Also, by Daniell-Kolmogorov

extension theorem, finite many points of distributions will define a stochastic process. And for $(E_n(x_1), E_n(x_2), ..., E_n(x_k))$ and $(B(x_1)mB(x_2), ..., B(x_k))$ will define the empirical process of uniform distribution and Brownian bridge respectively. The following □

$$\begin{array}{ccc} \begin{bmatrix} E_n(x_1) \\ \vdots \\ E_n(x_k) \end{bmatrix} & \xrightarrow{d} & \begin{bmatrix} B(x_1) \\ \vdots \\ B(x_k) \end{bmatrix} \\ \downarrow & & \downarrow \\ E_n & & B \end{array}$$

Figure 1: Proof flow of Theorem 6

Finally, let's put everything together. First, since $E_n \to B$, it's clearly that

$$\sup_{0\le x\le 1} |E_n(x)| \xrightarrow{d} \sup_{0\le x\le 1} |B(x)|$$

Namely, the Kolmogorov statistics of the uniform distribution on [0,1] will converge in distribution to the absolute supremum norm of a Brownian bridge.

As to general distribution F. Let G denotes the uniform distribution on [0, 1]. Consider the Kolmogorov statistics

$$\begin{aligned} D_n :&= \sqrt{n} \sup_{0\le x\le 1} |\hat{F}_n(x) - F(x)| \\ &= \sqrt{n} \sup_{0\le x\le 1} |G[\hat{F}_n(x)] - G[F(x)]| \\ &\xrightarrow{d} \sup_{0\le x\le 1} |B(F(x))| \end{aligned}$$

This is the Donsker Theorem, the uniform central limit theorem for empirical process.