Givenko-Cantelli theorem, proof and simulations

Pietro Colaguori 1936709

November 2023

1 Theorem

The Glivenko-Cantelli Theorem is a fundamental result in probability theory and mathematical statistics. It describes the convergence of empirical cumulative distribution functions (ECDFs) to the true cumulative distribution function (CDF) of a random variable.

Let's dive into the statement, let $X_1, X_2, ..., X_n$ i.i.d random variables with cumulative distribution function (CDF) $F(x)$. Let $\tilde{F}_n(x)$ be the ECDF based on n observations:

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

Where $I(\cdot)$ is the indicator function.The Glivenko-Cantelli states that as the sample size n goes to infinity, the ECDF $\tilde{F}_n(x)$ converges to the true CDF $F(x)$ with probability 1. Meaning that:

$$sup_x|\tilde{F}_n(x) - F(x)| \to^{a.s.} 0$$

Where "a.s." stands for "almost surely".

We can notice some significant points about this theorem, they are the following:

1. **Uniform Convergence**: The theorem implies that the convergence is uniform over the entire range of x. This is a stronger form of convergence compared to pointwise convergence, ensuring that the discrepancy between the empirical and true CDF becomes arbitrarily small across the entire space.

2. **Practical Implications**: The Glivenko-Cantelli Theorem has important implications for statistical inference. It provides theoretical support for using the empirical distribution function as an estimator for the true distribution function.

3. **Sample Size Requirements**: This theorem does not provide information about the range of convergence but rather claims to work correctly as $n \to \infty$.

2 Proof

Consider the continous random variable X. Let's fix the values $-\infty = x_0 < x_1 < ... < x_{m-1}, x_m = \infty$ such that

$$F(x_j) - F(x_{j-1}) = \frac{1}{m}, j \in [1, m]$$

Now, $\forall x \in \Re, \exists j \in \{1, ..., m\}$ s.t. $x \in [x_{j-1}, x_j]$. We then observe the following:

$$F_n(x) - F(x) \leq F_n(x_j) - F(x_{j-1}) = F_n(x_j) - F(x_j) + \frac{1}{m}$$

$$F_n(x) - F(x) \geq F(x_{j-1}) - F_n(x_j) = F_n(x_{j-1}) - F(x_{j-1}) - \frac{1}{m}$$

This implies:

$$||F_n - F||_\infty = sup_{x\in\Re}|F_n(x) - F(x)| \leq max_{j\in\{1,...,m\}}|F_n(x_j) - F(x_j)| + \frac{1}{m}$$

We proceed by making the following observations:

$$max_{j\in\{1,...,m\}}|F_n(x_j) - F(x_j)| \to 0$$

for the Law of Large Numbers (LLN). So, we can guarantee that for any $\epsilon > 0$ and any integer m s.t. $\frac{1}{m} < \epsilon$, we can find N s.t. $\forall n \geq N$ we have that

$$max_{j\in\{1,...,m\}}|F_n(x_j) - F(x_j)| \leq \epsilon - \frac{1}{m}$$

Combining with the previous result, this further implies that

$$||F_n - F||_\infty \leq \epsilon$$

which is the definition of almost sure convergence. □

3 Simulation

We can proceed by simulating this theorem using a Python script in which we ask the user to insert a certain number of samples and, given a fixed CDF, which in this case is the uniform CDF, we observe how, as we increase the number of samples the ECDF converges with the true CDF.

```python
import streamlit as st
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import uniform
```

```python

def glivenko_cantelli_simulation(sample_size):
    # Generate random samples from a uniform distribution
    samples = np.random.uniform(0, 1, sample_size)

    # Sort the samples
    sorted_samples = np.sort(samples)

    # Calculate the empirical cumulative distribution function (ECDF)
    ecdf = np.arange(1, sample_size + 1) / sample_size

    # Plot the true cumulative distribution function (CDF) and the ECDF
    plt.plot(sorted_samples, ecdf, label='Empirical CDF')
    plt.plot(sorted_samples, uniform.cdf(sorted_samples), label='True CDF', linestyle='--')

    plt.title('Glivenko-Cantelli Theorem Simulation')
    plt.xlabel('Value')
    plt.ylabel('Cumulative Probability')
    plt.legend()
    plt.grid(True)
    st.pyplot()

# Streamlit app
st.title('Glivenko-Cantelli Theorem Simulation')

# Sidebar for user input
sample_size = st.sidebar.number_input('Sample Size:', value=1000, step=100)

# Display simulation
if st.button('Run Simulation'):
    glivenko_cantelli_simulation(sample_size)
```

If the number of samples is quite low, e.g. 10 samples, it is clear how the ECDF does not match the true CDF, however, as a consequence of the Glivenko-Cantelli theorem, as we increase the number of samples, e.g. to 1000, the ECDF converges wit hthe CDF.

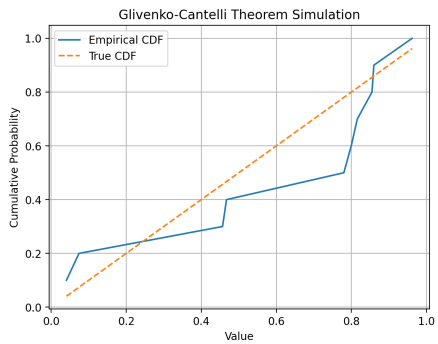

Figure 1: Simulation with number of samples = 10

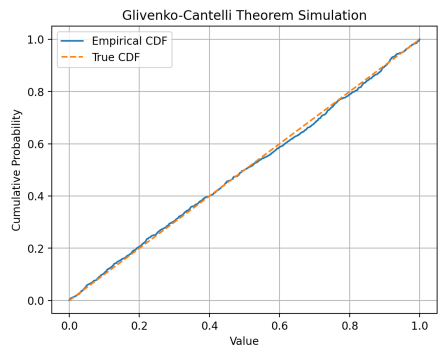

Figure 2: Simulation with number of samples = 1000