Glivenko-Cantelli Theorem

Chiara Iannicelli

November 2023

1 Introduction

The Glivenko-Cantelli theorem is a fundamental result in probability theory and mathematical statistics. It provides insights into the convergence of empirical distribution functions to the true underlying distribution as the sample size increases.

The theorem is named after mathematicians Dmitri Glivenko and Francesco Paolo Cantelli and is particularly important in the context of understanding how well the sample distribution function (empirical distribution function) approximates the true distribution function.

2 Statement

Assume that $X_1, X_2, ...$ are independent and identically distributed random variables in R with common cumulative distribution function $F(x)$. The empirical distribution function for $X_1, X_2, ..., X_n$ is defined by:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i,\infty)}(x) = \frac{1}{n} \left| \{ i | X_i \leq x, 1 \leq i \leq n \} \right|$$

where I_C is the indicator function of the set C. For every (fixed) $x, F_n(x)$ is a sequence of random variables which converge to $F(x)$ almost surely by the strong law of large numbers. Glivenko and Cantelli strengthened this result by proving uniform convergence of F_n to F.

3 Theorem

The theorem states as follows:

$$\|F_n - F\|_\infty = x \in Rsup \left| F_n(x) - F(x) \right| \longrightarrow 0 \quad \mathit{almost\,surely}$$

This theorem originates with Valery Glivenko[4] and Francesco Cantelli,[5] in 1933.

This is an uniform law of large numers:

$$
\begin{aligned}
\|F_n - F\|_\infty &= \sup_x |F_n(x) - F(x)| \\
&= \sup_x |P_n(X \leq x) - P[X \leq x]| \\
&\xrightarrow{as} 0,
\end{aligned}
$$

Figure 1: uniformity

where P_n is the empirical distribution that assigns mass $\frac{1}{n}$ to each X_i.

The law of large numbers says that, far all $x, P_n(X \leq x) \longrightarrow P(X \leq x)$. The GC Theorem says that this happens uniformly over x.

4 Simulation

An example of code the simulate the behaviour of the Glivenko-Cantelli Theorem is the following:

```
import matplotlib.pyplot as plt
import numpy as np

# True distribution function (CDF) for a uniform distribution
def true_distribution(x):
    return np.where(x < 0, 0, np.where(x <= 1, x, 1))

# Generate random samples from a uniform distribution
np.random.seed(42)  # Set seed for reproducibility
sample_size = 1000
samples = np.random.uniform(0, 1, sample_size)

# Calculate the empirical distribution function (EDF)
def empirical_distribution(sample, x):
    return np.sum(sample <= x) / len(sample)

# Calculate true distribution values
x_values = np.linspace(0, 1, 1000)
true_values = true_distribution(x_values)

# Calculate EDF values
edf_values = [empirical_distribution(samples, x) for x in x_values]
```

```python

# Plot the true distribution and EDF
plt.plot(x_values, true_values, label='True Distribution')
plt.step(x_values, edf_values, label='Empirical Distribution
    Function', where='post')
plt.xlabel('X-axis')
plt.ylabel('Cumulative Probability')
plt.title('Simulation of Glivenko-Cantelli Theorem')
plt.legend()
plt.show()

```

This generates the following graphic:

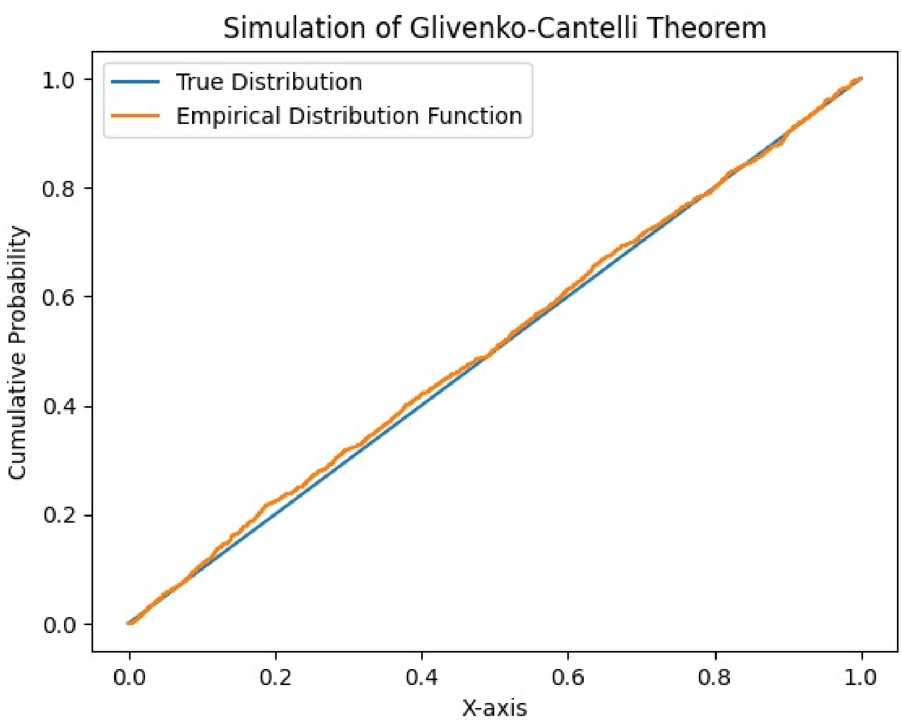

Figure 2: From colab